HindIdiom-Bench: Probing Declarative and Procedural Hindi Idiom Knowledge in LLMs

Anonymous ACL submission

Abstract

Idioms have been a stubborn challenge in the field of NLP: they are non-compositional, culturally grounded, and under-represented in pretraining data. Yet no benchmark exists to systematically evaluate LLM performance on Hindi idioms, leaving this gap unmeasured. So, we introduce **HindIdiom-Bench**, a benchmark to evaluate Hindi idiom knowledge in LLMs. It is designed to distinguish between knowing *what* an idiom means (declarative) and knowing *when* to use it correctly (procedural) since conflating them masks important gaps. We also introduce **SPICE** (Span Detection and Idiomatic Context Extraction), a LoRA based fine-tuning approach, to probe the effects of auxiliary learning. Our main findings show that models struggle on procedural tasks more than declarative tasks, as Task 2 lags Task 1 by 20.9 pp on average across model families; context helps models in disambiguating the correct meaning of an idiom; fine-tuning models on span detection and idiomatic context extraction improves their performance on higher level tasks despite using a disjoint set of idioms. This also shows that models rely on both internal information retrieval and reasoning, so idiom understanding can be improved by low-resource training as well.[1]

1 Introduction

Idioms are non-compositional multi-word expressions whose figurative meanings cannot be reliably derived from their constituents (Cheng and Bhat, 2024). Multi-word expressions have long posed a challenge for NLP due to their non-compositional nature and statistically low frequency (Constant et al. (2017); Sag et al. (2002)). Understanding of idiomatic use of phrases remains challenging for large language models (Chakrabarty et al. (2022); Haagsma et al. (2020); Phelps et al. (2024)). Furthermore, existing benchmarks that test the models on idiom understanding cater to two distinct demands: identifying what an idiom *means* (Kim et al., 2025; Madabushi et al., 2022) and judging when an idiom is *appropriately deployed* (Fu et al., 2025; Attia et al., 2026). The former tests whether models can recover an idiom's figurative meaning while the latter tests whether models can judge the correct use of an idiom. However, to the best of our knowledge, no benchmark of either type exists for Hindi, and no benchmark combines both for Hindi. This is a major gap as Hindi is the third most spoken language in the world (Ethnologue, 2026) yet under-represented in LLM pretraining (Singh et al., 2024). Most of the existing resources are either English centric or bilingual (Agrawal et al. (2018); Madabushi et al. (2022)). A richly annotated resource is needed to fill this gap with native Hindi idioms from modern and literary sources. Through this paper, we aim to bridge this gap in idiomatic understanding of models for Hindi.

Contributions.

1. **HindIdiom-Bench**: A Hindi idiom benchmark. It comprises two tasks that isolate declarative and procedural knowledge in the sense of Ryle and Tanney (2009). Task 1 tests declarative knowledge through meaning selection and Task 2 tests procedural knowledge as the model has to apply the knowledge in selecting the appropriate sentence for the idiom. The tasks are designed specifically to surface a robust empirical pattern:

[1] The final data and source code for experiments is available at https://anonymous.4open.science/r/hindidiom-B00C/

the declarative-procedural gap is wide for Hindi and persists across model scales and families.

2. **SPICE**: A LoRA fine-tuning approach that uses span detection in idiomatic/literal context without any MCQ signal. It improves performance on both benchmark tasks for mid-tier models, evidence that the knowledge gap for idioms can be slightly reduced through auxiliary learning.

3. **Systematic evaluation** of 16 LLMs revealing: (i) Task 2 consistently lags Task 1 across families and sizes; (ii) context helps Task 1 for smaller models but saturates near ceiling for large models; (iii) SPICE gains are model-size dependent, peaking at 4B-14B parameters. However, the gains reduce or reverse on Task 1 for larger models that already achieve a high score.

2 Related Work

Idiom Evaluation Dataset A variety of existing idiom corpora and benchmarks focus mostly on English and European languages (Flor et al. (2025); Moussallem et al. (2018); Chakrabarty et al. (2022); Haagsma et al. (2020)). Other than the language disparity, the findings of the existing works demonstrate a question- how to evaluate memorisation vs grounding. Cheng and Bhat (2024) created a dataset to test idiomatic expression reasoning scenario with context and without context. Madabushi et al. (2022) introduced a text based multilingual idiom evaluation methodology for English, Portuguese, and Galician as binary classification. Pickard et al. (2025) extended idiom evaluation to a multimodal ranking setting that is an analogue of procedural knowledge testing in language models. Several works on other languages (Kim et al. (2025); Liu et al. (2024); Fu et al. (2025); Attia et al. (2026)) create idiom benchmarks that demonstrate high accuracy on meaning selection can reflect memorisation rather than comprehension. These works motivate separating declarative from procedural knowledge.

Idiom Dataset for Hindi Agrawal et al. (2018) introduced the IMIL corpus for translating English idioms to Hindi, but contains no Hindi idiomatic contexts. Most of the existing corpora in Hindi and other Indo-Aryan languages focuses on Multi-Word Expressions and POS-tagging (Singh et al. (2016); Jain and Vaidya (2024); Sankaran et al. (2008); Vyas et al. (2014); Singh et al. (2018). These may include some idioms but it is minimal compared to the size of the dataset, for example, Parseme contains 12 VIDs (Verbal Idiomatic Expressions) out of 399 VMWEs (Verbal Multi-Word Expressions). Works like these, have contributed to the growth of Hindi NLP datasets in recent years (Singh et al. (2024)).

Auxiliary tasks, LoRA, and contextual idiom representation. Auxiliary-task learning has a long history in NLP (Ruder, 2017). Hu et al. (2022) showed that LoRA enables low-cost task-specific adaptation of large models, making auxiliary interventions practical. Cheng and Bhat (2024) showed that context removal sharply changes idiom interpretation, motivating contrastive idiomatic/literal data as a representation-learning signal. Based on the evidence from such studies, we aim to test a new LoRA (Low-Rank Adaptation) based fine-tuning approach, SPICE, that trains the model on span annotation and identification of idiomatic or literal usage of the idiom in a given sentence.

3 Data Resources

Two disjoint datasets underlie this work: the **HindIdiom-Bench evaluation benchmark** (Task 1 + Task 2), and a separate Hindi idiom fine-tuning corpus for **SPICE**. The two are constructed independently from different sources and contain no overlapping idioms, so any auxiliary intervention trained on the corpus is evaluated entirely out-of-distribution. Within the evaluation benchmark, the same 122 Hindi idioms underlie both Task 1 and Task 2.

3.1 Evaluation Benchmark: Task 1 (Meaning Selection)

Source and selection. Task 1 covers the shared 122 Hindi idioms sourced from

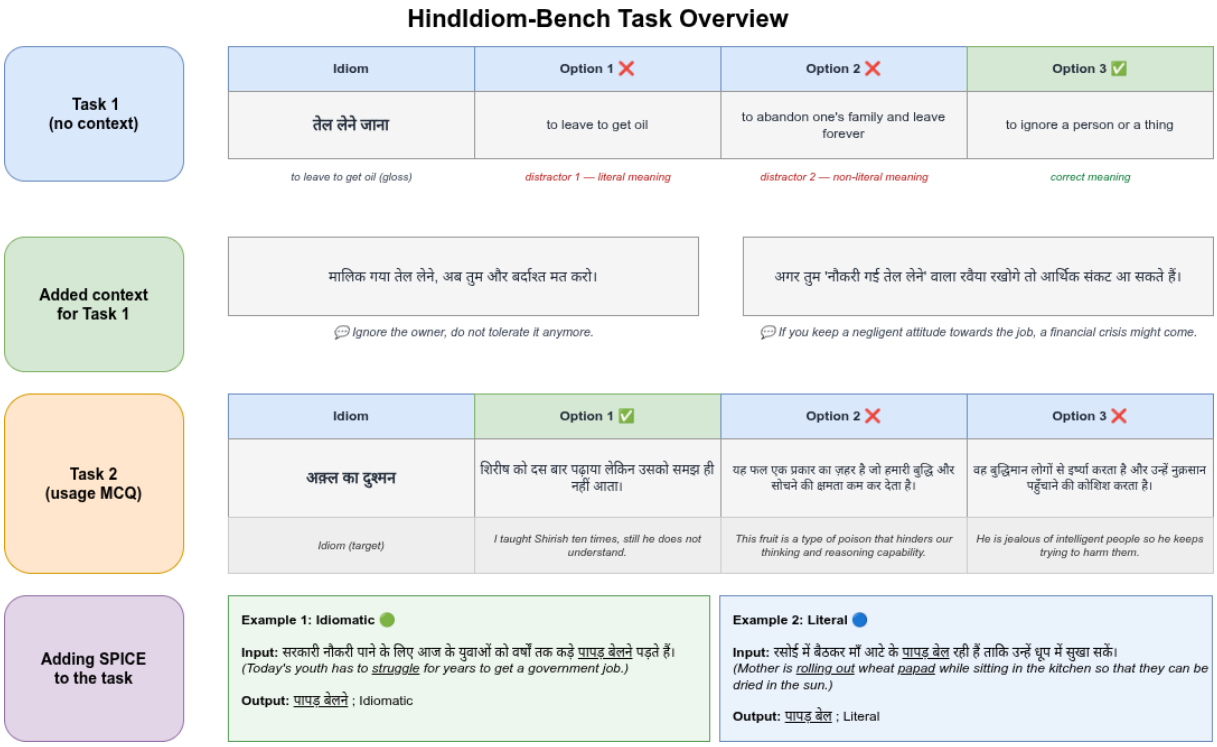

Figure 1: HindIdiom-Bench task overview.

Wiktionary (https://en.wiktionary.org/wiki/Category:Hindi_idiom), inspired by works like Kim et al. (2025). Idioms were excluded if they contained offensive language, were highly compositional[2], or appeared in the fine-tuning corpus. The quality-over-quantity scale matches recent practice in expert-annotated idiom benchmarks (Kim et al., 2025; Fu et al., 2025).

MCQ construction. Each idiom has a 3-way MCQ: the correct idiomatic meaning (taken verbatim from its Wiktionary entry) and two distractors- a *semantically proximate* distractor, similar to the idiomatic meaning but closer to the literal meaning of constituent words[3], and a *literal translation of the idiom*. The rationale behind the distractors is to trick a model relying on surface pattern matching rather than genuine comprehension. Because correct meanings are sourced from Wiktionary, the labelling step is essentially objective and does not require additional human inter-annotator agreement (see §3.3).

Context conditions. Each idiom is evaluated in context and no context conditions. In no context condition, only the phrase is given with the options, forcing reliance on pre-training alone. For adding context, two human-written sentences containing the idiom phrase are given:

- **Simple context** (Sentence 1): a naturalistic short passage supporting the figurative reading.
- **Complex context** (Sentence 2): a sentence introducing inferential difficulty including negation, conditional framing, or embedded clauses. Motivated by evidence that negation poses a systematic challenge for LLMs (Rezaei and Blanco, 2025), the complex condition increases the sentence load.

In the main results, the context condition presents both Sentence 1 and Sentence 2 together; the no-context condition presents the idiom alone. The joint presentation of simple and complex context sentences is intended to

[2]Highly compositional idioms included idioms for which the figurative meaning was highly derivable from the constituents. For example, '8 peher', where peher is a unit of time in Hindi. This idiom translates into 24-hours and the figurative meaning is also one whole day.

[3]For example, if the literal translation of an idiom is 'show mirror', its semantically proximate distractor would be to help someone put on makeup

reflect naturalistic discourse conditions; future work can decompose these effects.

3.2 Evaluation Benchmark: Task 2 (Sentence Selection)

Task 2 covers 122 Hindi idioms. Each item presents the idiom together with three candidate sentences in Hindi that represent a situation where the idiom can be used. The sentences do not have the idiom phrase in them, so the model has to identify the sentence where the idiom can be applied. This task isolates *procedural* knowledge: not what the idiom means, but where to use it correctly.

Sentence construction. For each idiom, the annotators wrote one correct sentence describing a situation where the idiom can be correctly used, and two distractor sentences:

- **Literal distractor**: A situation which aligns with the compositional meaning of the idiom phrase[4].
- **Semantically Proximate distractor**: A situation where idiomatic use seems more plausible than the correct idiomatic situation as it is closer to the compositional meaning of the idiom.

3.3 Annotation Quality and Inter-Annotator Agreement

As the sentences for Task 2 are written by the author, an inter-annotator agreement was computed over a stratified 20-idiom sample (16% of the 122-item benchmark), reviewed independently by four reviewers- three of them are experts in Hindi linguistics, and one on Multi-Word Expressions. This sample size matches Artstein and Poesio (2008)'s recommendation for expert annotation tasks where the goal is validating the annotation procedure rather than estimating a population-level disagreement rate.

Pairwise Cohen's κ ranged from 0.8953 to 1.0000 (mean pairwise $\kappa = 0.9126$), indicating very high agreement. Notably, the MWE specialist achieved $\kappa = 1.0000$ against the consensus gold standard, while the three other linguists achieved $\kappa = 0.8953$, 0.9475, and 0.8953 (Table 1, Figure 2). This shows that even highly qualified Hindi experts disagree on a small but non-trivial fraction of Task 2 items, while a domain specialist with prior exposure to MWE data collection achieves perfect alignment. This pattern supports our claim that Task 2 requires embedded familiarity with the idiom usage (Pavlick and Kwiatkowski (2019)).

Reviewer	κ vs. gold
Annotator 1 (Hindi expert)	0.8953
Annotator 2 (Hindi expert)	0.9475
Annotator 3 (MWE specialist)	1.0000
Annotator 4 (Hindi expert)	0.8953
Mean pairwise κ	0.9126

Table 1: Task 2 inter-annotator agreement. Cohen's κ for each of four reviewers against the consensus gold label. The MWE-specialist alignment of $\kappa = 1.0$ supports the difficulty interpretation.

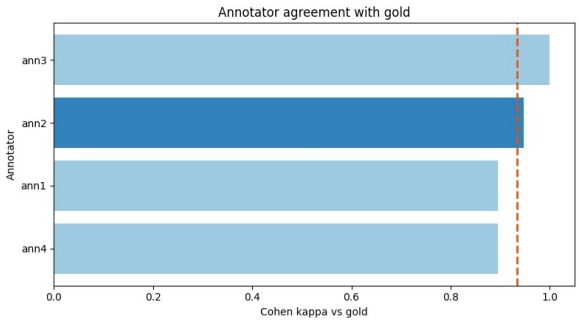

Figure 2: Task 2 annotator agreement with gold. The dashed line marks the mean pairwise κ (0.9126). The MWE specialist (ann3) achieves perfect alignment.

3.4 Fine-Tuning Corpus

The fine-tuning corpus is a separately collected resource used in this paper for SPICE. It is constructed to be disjoint from the evaluation benchmark.

Source. 456 Hindi idioms drawn from the Class 9 Hindi grammar textbook *Hindi Vyakaran & Rachna*, published by the Board of Secondary Education, Rajasthan (BSER) and freely distributed by the Government of Rajasthan for educational use[5]. The textbook

[4]For example, if the idiom translates into "sleep after selling all the horses", this distractor sentence would translate into "That person did not have a place to sleep, so he sold his horses and slept in the stable." The idiom actually means to sleep peacefully after getting rid of all problems.

[5]https://rajeduboard.rajasthan.gov.in/books-2019/cls9/Hindi%20Vyakaran%20&%20Rachna.pdf

Sentence configuration	Idioms
2 idiomatic + 2 literal	290
2 idiomatic + 1 literal	48
2 idiomatic + 0 literal	118
Total	456

Table 2: Fine-tuning corpus: per-idiom sentence configuration after human review. Literal sentences received stricter filtering, with several removed for low naturalness (a known difficulty for highly opaque idioms).

represents a curated, pedagogically validated source of standard Hindi idiomatic expressions and is independent of Wiktionary, the Task 1 source. All 456 idioms are confirmed disjoint from the evaluation benchmark by exact string matching after Unicode (NFC) normalization.

Sentence generation. For each idiom, up to four sentences were generated through the Model-in-the-Loop (MILO) paradigm (Wang et al., 2024; Chakrabarty et al., 2022), an LLM proposes candidate sentences and human annotators make the final accept/edit/reject decision. Each sentence was validated by native Hindi speakers. For most idioms, a literal usage was possible, so two idiomatic usage sentences and two literal usage sentences were created. However, to maintain the quality, some literal usage sentences were either rejected by Hindi speakers (did not sound natural or plausible) or not created by us (for idioms where literal usage sounds impossible). The final per-idiom breakdown is shown in Table 2.

Human review. All generated sentences were reviewed by 15 native Hindi speakers using an approve/edit/reject protocol, evaluating naturalness, grammatical correctness, and sense disambiguation.

Train/test split and contrastive structuring. The corpus is split 80/20 by phrase where no idiom appears in both splits. Same-phrase idiomatic-literal pairs are interleaved in the training stream to place contrastive examples in adjacent positions before each epoch shuffle. Literal sentences are oversampled 1.5× in the train split as harder negatives. Final split statistics are in Table 3.

	Train	Test
Unique idioms	359	90
Idiomatic sentences	718	180
Literal sentences	745	124
Total records (post-oversample)	1463	304

Table 3: SPICE fine-tuning corpus split. 80/20 by phrase, disjoint from the evaluation benchmark by construction.

4 SPICE: Span Prediction for Idiomatic Context Extraction

SPICE is a LoRA fine-tuning approach that trains a model on the span detection of idiom and idiomatic or literal sense detection. It evaluates zero-shot transfer to the HindIdiomBench MCQ tasks. No MCQ signal is seen during training. SPICE therefore probes whether span-level sense discrimination is a transferable representation that aids the model in further disambiguating the meaning and usage of an idiom.

Auxiliary objective. Given a Hindi sentence the model is trained to output:

- Idiomatic: <phrase>: if the idiom is used figuratively, or
- None: if the idiom is used literally.

Loss is applied only on completion tokens (prompt tokens masked with -100), so the model optimizes only its label prediction. This is structurally simpler than the MCQ tasks that the benchmark evaluates.

Contrastive signal. Same-phrase idiomatic/literal pairs are interleaved during training, so contrastive examples appear in adjacent positions before each epoch shuffle. Literal sentences are oversampled 1.5× as harder negatives. The design forces the model to attend to contextual cues distinguishing figurative from literal usage rather than relying on lexical surface features of the phrase.

LoRA configuration. LoRA (Hu et al., 2022) rank is scaled inversely with model size to constrain catastrophic forgetting (McCloskey and Cohen, 1989) on near-ceiling large models: r=16 for <5B, r=8 for 5-20B, r=4 for ≥27B. We set $\alpha = 2r$ throughout. Full hyperparameters are reported in Appendix B.

Span-detection training quality. Across all 16 models, held-out span-detection eval loss converged below 0.04 (best-epoch range 0.006-0.039; see Appendix C for per-model values). All trained models thus learned the auxiliary objective to high fidelity; subsequent MCQ-task variation reflects transfer behaviour rather than auxiliary-task failure.

Inference. At evaluation time, SPICE LoRA weights are merged into the base model. The evaluation prompt explicitly references the model's span-detection training history, activating the learned sense-discrimination representations for the MCQ task (Appendix A).

5 Experiments

Models. We evaluate 16 instruction-tuned LLMs: Qwen3-0.6B/1.7B/4B/8B/14B/32B, Llama-3.2-1B/3B and Llama-3.1-8B/70B, Gemma-3-4B/12B/27B, and Gemma-4-E2B/E4B/31B. All models are loaded in bfloat16 without quantization.

Conditions. Task 1 is evaluated under {base, SPICE} × {context, no-context}. Context condition has both Sentence 1 (standard) and Sentence 2 (complex) jointly in a single prompt. Task 2 is evaluated under {base, SPICE}. All Task 1 evaluation uses the decode method: the model generates a chain-of-thought response and the final answer label is extracted via regex; logit-scoring fallback is applied only when extraction fails.

Metric. Primary metric is accuracy. For Task 2 we additionally analyze the hard-versus-soft distractor error split.

6 Results

6.1 Main Results

Table 4 reports the main accuracy results.. Across the 16 models, the mean Task 1 (context, base) accuracy is 76.8% while Task 2 (base) is 55.9%, and a 20.9 pp average gap that is the central empirical finding of the benchmark.

6.2 Context Aids Meaning Selection

Context substantially improves Task 1 accuracy for smaller and mid-tier models, consistent with prior evidence that contextual sentences improve idiom understanding Cheng and Bhat (2024). For example, Qwen3-8B improves from 67.2% (no context) to 78.7% (with context), a gain of 11.5 pp. Near-ceiling models, however, show no benefit or a slight reversal: Llama-3.1-70B scores 95.08% without context and 94.26% with context. This pattern is consistent with Kim et al. (2025)'s observation that high-accuracy idiom models may be engaging in pretraining-based recall rather than contextual inference.

6.3 SPICE + Context Yields the Best Task 1 Performance

The SPICE + context condition achieves the highest Task 1 accuracy for mid-tier models. Qwen3-8B reaches 91.8% (+13.1 pp over base+context); Gemma-3-4B reaches 87.7% (+13.1 pp). Near-ceiling models (Gemma-4-31B: 99.2% base, 99.2% SPICE) show only marginal incremental or no gains. The gains are highest in the 4B-14B regime; smaller and larger models exhibit the size-dependent patterns.

6.4 Task 2 Consistently Lags Task 1

Task 2 lags Task 1 (context, base) by substantial margins across all models. Qwen3-8B: 78.7% (T1) vs. 46.7% (T2), a 32.0 pp gap. Gemma-3-4B: 74.6% vs. 34.4%, a 40.2 pp gap. Even the strongest model, Gemma-4-31B, shows a 5.7 pp gap (99.2% vs. 93.4%). The pattern replicates cross-linguistically the understanding-use gap documented for Arabic and English (Attia et al., 2026) and Chinese (Fu et al., 2025), and amplifies it for Hindi, consistent with lower pretraining exposure to Hindi idiomatic text. The mean Task 1→Task 2 gap (≈ 21 pp) is our primary operationalisation of the pragmatic-grounding hypothesis: models that score high on T1 but low on T2 demonstrate idiom knowledge that is declarative but not procedural.

6.5 SPICE Effect Is Model-Size Dependent

SPICE effects vary significantly with model size, consistent with our hypothesis. Mid-tier models (4B-14B) benefit most: Qwen3-8B gains +13.1 pp on T1 (context) and +10.7 pp on T2; Gemma-3-4B gains +13.1 pp and

Model	T1 Context		T1 No context		T2 Appropriateness	
	Base	SPICE	Base	SPICE	Base	SPICE
Qwen3						
0.6B	36.89	42.62	25.41	35.25	23.77	30.33
1.7B	54.92	66.39	46.72	59.02	34.43	44.26
4B	72.13	77.87	55.74	73.77	40.16	52.46
8B	78.69	91.80	67.21	85.25	46.72	57.38
14B	80.33	87.70	71.31	86.07	59.84	59.02
32B	91.80	96.72	85.25	90.98	67.21	67.21
Llama						
3.2-1B	36.89	33.61	25.41	31.15	27.05	31.97
3.2-3B	64.75	70.49	48.36	61.48	39.34	40.16
3.1-8B	77.05	88.52	60.66	79.51	40.98	44.26
3.1-70B	94.26	93.44	95.08	91.80	86.07	81.15
Gemma-3						
4B	74.59	87.70	67.21	81.15	34.43	50.00
12B	91.80	94.26	87.70	89.34	79.51	80.33
27B	95.90	96.72	95.08	93.44	85.25	88.52
Gemma-4						
E2B	83.61	91.80	79.51	81.15	58.20	59.02
E4B	94.26	97.54	91.80	95.08	77.05	81.15
31B	99.18	99.18	100.00	100.00	93.44	94.26
Mean	76.74	82.27	68.85	77.10	55.84	60.09

Table 4: Accuracy (%) on HindIdiom-Bench. T1 context presents Sentence 1 and Sentence 2 jointly; T1 no-context presents the idiom alone; T2 evaluates sentence appropriateness over the same idiom set. Chance = 33.3% for both tasks.

+15.6 pp. Near-ceiling large models show small or negative changes: Llama-3.1-70B drops from 86.0% to 81.1% on Task 2, suggesting that span-detection training at this scale partially disrupts existing pragmatic representations through catastrophic interference (McCloskey and Cohen, 1989). The smallest models (0.6B-1B) show small or inconsistent gains, reflecting limited capacity to encode the contrastive signal.

6.6 SPICE Effect on Task 2

SPICE improves Task 2 accuracy for 14 out of 16 models despite no Task 2 examples being seen during training. The largest gains again cluster in the mid-tier range: Gemma-3-4B +15.6 pp, Qwen3-4B +12.3 pp, Qwen3-8B +10.7 pp, Qwen3-1.7B +9.8 pp. This is the strongest single evidence that span-level context discrimination is a genuinely transferable representation: a model trained only to label the phrase and the idiomatic vs. literal usage in context acquires structure that supports not only meaning selection (T1) but also situation-based application of the idiom (T2). Importantly, SPICE is a single auxiliary objective among many that the benchmark can be used to evaluate; we read these gains as evidence that the procedural gap is targetable in principle, not as a claim that span detection is the uniquely correct intervention.

7 Analysis

7.1 Literal-Confusion Error Profile

A recurring failure mode is selecting the *literal* reading of an idiom over the other (non-literal but incorrect) distractor. For each error a model commits, we ask whether the chosen option was the literal one. Both tasks expose a literal foil: in Task 1 it is the literal-meaning gloss, and in Task 2 it is the literal-sense sentence (Scenario 2). Table 5 reports, for each model, the share of errors that fall on the literal option, under the base model and the SPICE-tuned model, for Task 1 (context), Task 1 (no-context), and Task 2. To keep the three tasks comparable, Task 1 is restricted to the 122 idioms shared with Task 2.

Effect of Context and SPICE Since the main issue this highlighted in this type of errors is that LLMs are unable to disambiguate between figurative and literal use of words in phrases. For this, we experiment adding con-

Model	Base			SPICE		
	T1	$T1_{nc}$	T2	T1	$T1_{nc}$	T2
Qwen3-0.6B	76.3	65.1	67.4	57.0	61.8	52.4
Qwen3-1.7B	65.7	63.5	46.8	48.7	47.8	26.9
Qwen3-4B	81.4	56.6	40.3	50.0	34.4	31.0
Qwen3-8B	61.2	35.9	43.8	38.1	36.8	48.1
Qwen3-14B	56.2	50.0	38.8	37.9	16.7	37.1
Qwen3-32B	57.9	33.3	37.5	44.4	33.3	32.7
Llama-3.2-1B	68.9	54.0	60.2	56.6	59.8	64.6
Llama-3.2-3B	81.4	72.6	34.2	71.8	68.1	64.2
Llama-3.1-8B	73.7	69.6	40.8	48.3	45.8	50.7
Llama-3.1-70B	58.3	83.3	17.6	76.5	22.2	34.3
Gemma-3-4B	88.7	78.6	67.1	70.0	70.8	49.2
Gemma-3-12B	75.0	50.0	12.0	53.8	23.1	47.2
Gemma-3-27B	63.6	33.3	38.9	37.5	25.0	11.5
Gemma-4-E2B	78.0	70.8	56.9	52.4	50.0	56.8
Gemma-4-E4B	57.1	60.0	50.0	28.6	33.3	49.2
Gemma-4-31B	100.0	-	12.5	100.0	-	22.2

Table 5: Literal-confusion share: percentage of a model's errors in which it picks the literal option rather than the other (non-literal) distractor. Reported for Task 1 (context), Task 1 no-context ($T1_{nc}$), and Task 2, under base and SPICE conditions. Task 1 is restricted to the 122 idioms shared with Task 2. "-" marks conditions with no errors to classify.

text (in TASK 1) and SPICE (in both tasks). In Task 1, it is observed that adding context lowers the share for most models, consistent with context steering models away from the literal reading. SPICE further improves performance as models are trained to detect a group of words denoting a single idiomatic meaning in certain context. SPICE+Context condition gives the best results in TASK 1, most visible on mid-sized models. In TASK 2 as well SPICE improves the performance.

7.2 Translation Trap

We also conducted a post-hoc qualitative analysis of model outputs. It is observed that English translation-based approach is misleading the models. LLMs attempt translation when confused, misleading to worse performance.

- **Small Models (<2B):** They often hallucinate nonsensical translations. For instance, "Prince of an egg" becomes "The child's strength is the father's strength". This happens when they attempt to translate the idiom word-by-word, and when they lack the exact translation of a word, they find a semantically closer word (like 'Dog' instead of 'Owl'). This leads to such semantically distant translations of the idiom which is neither close to the literal meaning nor actual figurative meaning.

- **Medium Models (4-10B):** In these models, translations are better and closer to the literal translation. This translation misleads them towards the literal or the semantically proximate distractor over the correct option. However, this helps them in Task 1 for some idioms where an English equivalent is also present, for example, 'Crocodile's tears' is an idiom in Hindi (Hindi translation of the idiom) and English both. Such idioms were intentionally kept to observe this translation effect.

- **Larger Models (>10B):** These models did not seem to rely on just translation. Even when they used translation for some idioms, it resulted in better clarity towards the right answer. However, they still struggled in Task 2.

7.3 Idiom Categorisation Based on Difficulty

- **Cultural or Region-Specific Origin:** Some idioms use terms that are commonly used either in a cultural context (like "conch shell" which holds religious significance in some cultures) or idioms that are common only in certain regions. In such cases, meaning disambiguation became difficult for models.

- **Sarcastic Role Reversal:** Certain idioms include a sarcastic role reversal. For instance, an idiom that translates to 'the thief scolds the police' has a sarcastic role reversal where the thief (the usual patient of the verb 'scolding') is acting as the agent to denote a figurative meaning 'culprit blaming other'. In such cases, models struggle a lot.

Limitations

Benchmark size. HindIdiom-Bench contains 122 idioms shared across Task 1 and Task 2, reflecting a quality-over-quantity design. Each item is individually constructed and verified. This limits statistical power for fine-grained subgroup analysis.

Model coverage. While we evaluate a diverse set of instruction-tuned LLMs across multiple model families and parameter scales, some frontier-scale LLMs above the 70B parameter range were omitted due to computational and resource constraints. However, the evaluated suite spans multiple architectures, parameter regimes, and training paradigms, allowing robust cross-family comparison.

Single auxiliary intervention. SPICE is one of many possible auxiliary objectives. Whether other auxiliary tasks (e.g., paraphrase generation, sense gloss reconstruction, contrastive embedding) transfer comparably or better is an open question the benchmark enables but does not answer here.

Ethics Statement

All annotators involved in sentence validation and review were fluent native Hindi speakers and participated voluntarily. The benchmark does not contain personally identifiable information. Idioms containing offensive or abusive language were excluded during dataset construction. We acknowledge that idiomatic interpretation may vary across dialects, sociolects, and regions of Hindi usage, and future versions of the benchmark may broaden dialectal coverage.

References

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Sharma. 2018. No more beating about the bush : A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Mena Attia, Aashiq Muhamed, Mai Alkhamissi, Thamar Solorio, and Mona Diab. 2026. Beyond understanding: Evaluating the pragmatic gap in LLMs' cultural processing of figurative language. In *Proceedings of the 2026 Conference of the European Chapter of the Association for Computational Linguistics*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations. *Preprint*, arXiv:2205.12404.

Fangzhou Cheng and Suma Bhat. 2024. No context needed: Contextual quandary in idiomatic reasoning with pre-trained language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Ethnologue. 2026. What are the top 200 most spoken languages? Accessed: 2026-05-24.

Michael Flor, Xinyi Liu, and Anna Feldman. 2025. A survey of idiom datasets for psycholinguistic and computational research. *Preprint*, arXiv:2508.11828.

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.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kanishka Jain and Ashwini Vaidya. 2024. Revisiting VMWEs in Hindi: Annotating layers of predication. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 98–105, Torino, Italia. ELRA and ICCL.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

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

Tayyar Madabushi, Harish, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation*.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

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

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. SemEval-2025 task 1: AdMIRe — advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation*.

MohammadHossein Rezaei and Eduardo Blanco. 2025. Making language models robust against negation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 8123–8142.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Gilbert Ryle and Julia Tanney. 2009. *The concept of mind*. Routledge.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Baskaran Sankaran, Kalika Bali, Monojit Choudhury, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha, and K.V. Subbarao. 2008. A common parts-of-speech tagset framework for Indian languages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Dhirendra Singh, Sudha Bhingardive, and Pushpak Bhattacharyya. 2016. Multiword expressions dataset for Indian languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2331–2335, Portorož, Slovenia. European Language Resources Association (ELRA).

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha" Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. A twitter corpus for hindi-english code mixed pos tagging. In *Proceedings of the sixth international workshop on natural language processing for social media*, pages 12–17.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 974–979.

Yifan Wang, David Stevens, Pranay Shah, Wenwen Jiang, Miao Liu, Xu Chen, Robert Kuo, Na Li, Boying Gong, Daniel Lee, Jiabo Hu, Ning Zhang, and Bob Kamma. 2024. Model-in-the-loop (MILO): Accelerating multimodal AI data annotation with LLMs. In *arXiv preprint arXiv:2409.10702*.

A System Prompts

All prompts are bilingual (Hindi + English). Base prompts are in §A.1; SPICE-specific prompts that reference span-detection training are in §A.2.

A.1 Base Evaluation Prompts

Task 1 — Context (base):

नीचे एक हिंदी मुहावरा और उसे इस्तेमाल करने वाला एक वाक्य दिया गया है तीन अर्थों में से सही अर्थ चुनिए जो वाक्य के संदर्भ में उचित हो

Below is a Hindi idiom and a sentence that uses it. Three possible meanings are given

and only one correctly captures what the idiom is expressing in this sentence. Read the sentence carefully to understand the context, then select the meaning that best fits.

Think briefly about which option fits the context, then end your response with exactly: Answer: <X> where <X> is A, B, or C.

Task 1 — No context (base):

नीचे एक हिंदी मुहावरा दिया गया है तीन अर्थों में से उस मुहावरे का सही लाक्षणिक अर्थ चुनिए

Below is a Hindi idiom. Three possible meanings are given only one correctly describes the figurative (idiomatic) meaning of this phrase, i.e., what it conventionally expresses when used as an idiom rather than literally.

Think briefly about the conventional meaning of this idiom, then end your response with exactly: Answer: <X> where <X> is A, B, or C.

Task 2 (base):

नीचे एक हिंदी मुहावरा और तीन वाक्य दिए गए हैं इनमें से केवल एक वाक्य मुहावरे का सही और उचित प्रयोग करता है अर्थात् मुहावरे का लाक्षणिक अर्थ संदर्भ के अनुसार सटीक बैठता है बाकी दो वाक्य या तो मुहावरे का शाब्दिक अर्थ लेते हैं या विषय से हटे हुए हैं

Below is a Hindi idiom and three sentences. Only ONE sentence uses the idiom correctly in its figurative (idiomatic) sense the phrase carries its culturally understood meaning and fits naturally in context. The other two either take the words literally (as if they mean exactly what they say) or use the phrase in an unrelated way.

Think briefly about each option, then end your response with exactly: Answer: <X> where <X> is A, B, or C.

A.2 SPICE Evaluation Prompts

Task 1 — Context (SPICE):

आपको हिंदी मुहावरों की लाक्षणिक पहचान का विशेष प्रशिक्षण दिया गया है आप जानते हैं कि किसी वाक्य में मुहावरा अपने लाक्षणिक अर्थ में प्रयुक्त हो रहा है या शाब्दिक अर्थ में

You have been specially trained to detect whether a Hindi idiom is being used figuratively (idiomatic) or literally in a sentence you know how to read context and identify when a phrase carries its deeper, culturally understood meaning.

The sentence below uses the given idiom in its FIGURATIVE sense the kind of usage you were trained to recognise and extract as an idiomatic span. Now go one step further: identify which of the three meanings best captures what this figurative usage is expressing in this particular context.

Think about what the idiom is communicating figuratively, then end your response with exactly: Answer: <X> where <X> is A, B, or C.

Task 1 — No context (SPICE):

आपको हिंदी मुहावरों की लाक्षणिक पहचान का विशेष प्रशिक्षण दिया गया है

You have been specially trained on the figurative (idiomatic) meanings of Hindi idioms you have seen these phrases used in sentences and learned to distinguish their idiomatic meaning from their literal surface meaning.

Below is a Hindi idiom. From your training, you have encountered this phrase in idiomatic contexts and learned what it conventionally expresses. Select the option that correctly describes its FIGURATIVE meaning.

Think about what this idiom conventionally expresses from your training, then end your response with exactly: Answer: <X> where <X> is A, B, or C.

Task 2 (SPICE):

आपको हिंदी मुहावरों की लाक्षणिक पहचान का विशेष प्रशिक्षण दिया गया है आप जानते हैं कि किसी वाक्य में मुहावरा अपने लाक्षणिक अर्थ में प्रयुक्त हो रहा है या शाब्दिक अर्थ में

You have been specially trained to detect whether a Hindi idiom is used in its FIGURATIVE (idiomatic) sense or its LITERAL sense in a given sentence you know how to read context and identify when a phrase carries its deeper, culturally understood meaning versus when it is taken at face value.

Apply that exact skill here: three sentences are given, each containing the idiom. For each one, ask yourself 'Is this idiom being used figuratively here, the way I was trained to detect? Would I output Idiomatic: or None for this sentence?' Exactly ONE option would give you an idiomatic span that is your answer.

Think through each option carefully, then end your response with exactly: Answer: <X> where <X> is A, B, or C.

A.3 SPICE Training Prompt (Span Detection)

नीचे एक हिंदी वाक्य और एक मुहावरा दिया गया है आपको यह निर्धारित करना है कि इस वाक्य में मुहावरे का प्रयोग लाक्षणिक (idiomatic) रूप में हो रहा है या शाब्दिक (literal) रूप में

Below is a Hindi sentence and a Hindi idiom. Your task is to determine whether the idiom is used in its FIGURATIVE (idiomatic) sense or its LITERAL sense.

Key distinction: FIGURATIVE the phrase carries a culturally understood meaning beyond what its individual words literally say. LITERAL the words are used at face value.

Output rules: If FIGURATIVE → output exactly: Idiomatic: <idiom phrase>. If LITERAL → output exactly: None.

Hyperparameter	Value
Optimizer	AdamW 8-bit
Learning rate	2×10^{-4}
LR scheduler	Cosine
Warmup ratio	0.05
Weight decay	0.01
Batch size	2 (per device)
Gradient accumulation	8
Epochs	3 (2 for $\geq$27B models)
LoRA r	16 (<5B), 8 (5-20B), 4 ($\geq$27B)
LoRA α	$2r$
LoRA dropout	0.05
Precision	bfloat16
Max sequence length	1024

Table 6: SPICE training hyperparameters.

B SPICE Hyperparameters

C Span-Detection Eval Loss per Model

Model	Best eval loss
Qwen3-0.6B	0.0232
Qwen3-1.7B	0.0394
Qwen3-4B	0.0077
Qwen3-8B	0.0065
Qwen3-14B	0.0091
Qwen3-32B	0.0079
Llama-3.2-1B	0.0223
Llama-3.2-3B	0.0232
Llama-3.1-8B	0.0140
Llama-3.1-70B	0.0111
Gemma-3-4B	0.0158
Gemma-3-12B	0.0105
Gemma-3-27B	0.0147
Gemma-4-E2B	0.0141
Gemma-4-E4B	0.0132
Gemma-4-31B	0.0113

Table 7: Best held-out span-detection eval loss across SPICE training epochs. All trained models converged below 0.04, confirming reliable acquisition of the auxiliary objective.