ID10M-JAM

Stress-testing Idiomatic Identification In Challenging Contexts

Kai Golan Hashiloni, Lior Livyatan, Ofri Hefetz,
Alon Mannor, Bar Cohen, Kfir Bar

Reichman University

DEFINITIONS

Idioms = an expression whose meaning cannot be inferred from the combined meanings of its parts.

Potentially Idiomatic Expressions (PIEs) = Expressions that may be an idiom or not, **depending** on the context

MOTIVATION

Why Idioms Matter

- Same surface form → literal / figurative meaning
- Errors cause downstream failures:
  - Machine Translation
  - Semantic analysis

Key challenge: Context is necessary — but what happens if it is misleading?

RESEARCH QUESTION

Context may contain adversarial cues that can bias models toward the wrong interpretation of PIEs

Are LLMs robust to misleading context?

ID10M → ID10M-JAM

Starting Point: ID10M [Tedeschi et. al. 2022]

Our extension: Create 3 hard variants per original sentence with adversarial context

Build a hard variant:

1. Keep the original sentence intact
2. Prepend a coherent context
3. Context biases toward the opposite meaning
   a. Idiomatic usage → add Literal cues
   b. literal usage → add Idiomatic cues

Crucial constraint: Still unambiguous for humans

Original sentence from ID10M:
They asked me if we can *go dutch*.

Same sentence with enriched context (hard variant):
We were eating some German bread with Belgian chocolate when they asked me if we can *go dutch*.

EVALUATION

1. A model (M) is correct on some **original sentences**
2. SM is the number of their associated **hard variants** (3X original sentences)
3. FM (flipped) is the part of SM, the model changed its prediction on
4. We define the negative drift: $ND_M = \frac{F_M}{S_M}$

		English		German	
Setting	**Model**	S_M	$ND_M(\%)\downarrow$	S_M	$ND_M(\%)\downarrow$
Zero-shot	GPT-4o mini	471.00±7.94	5.16±0.95	257.00±4.58	5.84±0.43
	Qwen2.5-72B	452.00±1.73	1.70±0.34	361.00±1.73	9.80±5.11
	Llama 4 Scout	481.00±6.93	6.30±0.76	292.00±6.24	8.89±1.43
	GPT-4o	483.00	6.42	333.00	12.01
	Claude 4 Sonnet	486.00	6.58	369.00	2.98
	Gemini 2.5 Flash	492.00±5.20	9.08±0.75	295.00±1.73	9.60±0.73
	Gemini 2.5 Pro	501.00	7.78	381.00	3.15
Few-shot +SC+CoTBest	GPT-4o mini	483.00±13.75	6.19±1.15	321.00±9.00	11.30±1.08
	Qwen2.5-72B	458.00±1.73	0.95±0.26	380.00±1.73	6.76±1.10
	Llama 4 Scout	487.00±3.46	6.36±0.31	308.00±13.86	8.74±1.00
	GPT-4o	495.00	6.06	390.00	8.72
	Claude 4 Sonnet	498.00	8.43	393.00	7.12
	Gemini 2.5 Flash	484.00±4.58	8.27±1.02	357.00±3.00	7.75±1.19
	Gemini 2.5 Pro	504.00	10.12	390.00	4.87
Reasoning LLMs	DeepSeek-R1	462.00	3.90	378.00	4.76
	o3-mini	477.00	5.45	324.00	8.95
Encoders	mBERT	298.80±6.91	23.96±3.66	206.40±20.28	19.78±1.86
	BERT	357.00±6.00	11.88±0.50	-	-
	GBERT	-	-	313.20±3.42	9.18±1.36

CONCLUATIONS

- We reveal systematic vulnerabilities in LLMs
- Robustness is **decoupled** from scale, reasoning, prompting sophistication

ACKNOWLEDGMENTS

This study is supported in part by the European Research Council (Intellexus, Project No.\@ 101118558).Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Unionor the European Research Council Executive Agency. Neither theEuropean Unionnor the granting authorities can be held responsible for them.

erc
ACL 2026 SAN DIEGO JULY 2-7
INTELLEXUS