Probing Slang Ambiguity in Large Language Models

Uri Kialy
sheffi1414@gmail.com
Arial University

Avi Shtarkberg
avishb1213@gmail.com
Arial University

Advanced Topics in LLM, Spring 2025
July 15, 2025

Abstract

Slang-informal words whose meanings drift rapidly and often deviate from classic lexicons-poses a unique challenge for Large Language Models (LLMs). While state of the art models excel at standard language understanding, their behaviour on slang remains mainly under-explored. We present a systematic investigation into how representative LLM families (BERT, GPT, Llama-2,and deepseek) interpret ambiguous slang terms. Leveraging the **OpenSubtitles** dialogue corpus and a novel **Filtered–Slang** subset aligned to model vocabularies, we apply *Testing with Concept Activation Vectors* (TCAV) to quantify model sensitivity to slang concepts and correlate these scores with prompt-based disambiguation accuracy. Our study reveals that: (i) model size correlates with slang competence; (ii) TCAV provides a reliable early-diagnostic signal of downstream slang performance; and (iii) models exhibit distinct layer-wise patterns for slang representation development, with peak discrimination occurring in middle layers rather than final layers, yet these temporal differences do not correlate with overall performance. We release all code and evaluation here: https://github.com/UriKialy/Probing-Slang-Ambiguity-in-LLM.git.

1 Introduction

Slang is an ever-evolving linguistic phenomenon that plays a central role in informal communication, online discourse, and popular culture [1]. In contrast to standard lexical items, slang terms are often polysemous and context dependent (e.g., "*drip*" for clothing style versus fluid). For safety-critical deployments such as content moderation or conversational agents, failure to resolve slang ambiguity can yield misinterpretations or biased outputs. Despite the surge of work on LLM evaluation, dedicated analyses of slang remain scarce.

Our research asks: *How effectively do current LLMs handle slang ambiguity, and what internal representations underpin their behaviour?* We tackle this question by combining interpretability (TCAV) with controlled evaluation datasets derived from film subtitles, providing both introspective and extrinsic perspectives.

2 Related Work

Early work examined lexical change on social media [2]; more recently, ElSherief et al. [3] studied bias in emergent slang. The main last research Sun et al. [4] has made significant changes. First, they created a huge dataset, Open-Sub, of 25,000 sentences , about 7,500 of them anointed as slang ,but manually, meaning they used human annotators to give their view on a sentence in a confidence level between 1 and 3. This huge dataset was novel in a way that the existing ones prior to that didn't have all of the following: Slang-containing sentences, non-slang sentences, Word-level sentences, Community of emergence, Time of emergence and

publicly accessible. Secondly, they tested their opensub dataset on state of the art models and saw good results. We saw this as an opportunity to take this testing to the next level. Concurrently, interpretability methods such as TCAV [5] have been adopted to audit many areas in deep learning . However, TCAV has not been used to probe slang concepts, nor has it been validated against behavioural metrics in LLMs. Our study bridges this gap.

3 Datasets

3.1 OpenSubtitles

In Sun et al. [4], the authors introduce a new dataset drawn from 100 English movies in the OpenSubtitles4 corpus, capturing slang usage contexts, movie metadata (US/UK region and production year), and word-level literal paraphrases. with the help of Amazon Mechanical Turk sampled and annotated 7,488 slang-containing sentences (3,583 unique terms), of which 836 were refined to include definitions and paraphrases after quality control. They also contribute 17,512 non-slang subtitle sentences unanimously agreed upon by annotators to serve as robust negative samples for slang detection. They leverage the natural conversational style and multilingual potential of OpenSubtitles to diversify beyond dictionary examples and enable future multilingual extensions.

3.2 Filtered–Slang (FS)

As we wanted to explore how LLM detects and evaluates ambiguous slang, we manually created a list of 140 words that appear in regular vocabulary such as 'fire', 'drip', 'wet', etc. The main idea was to prevent the models from understanding that the word isn't correct so the sentence must be slang, and moreover to learn how they deal with the ambiguity of correct words that is used in a slang sentence, perhaps it will lower the accuracy. The filtered dataset; contained 1,424 slang sentences and 2193 non-slang sentences.

4 Evaluation

We report results in three successive stages that mirror the amount of task-specific supervision each model receives.

4.1 Stage 1 – Zero-shot MNLI inference

We begin by treating slang detection as an NLI problem: each model must decide which of two hypotheses is more related to the premise sentence, *"This sentence contains slang."* or *"This sentence does not contain slang."*. The checkpoints evaluated are (i) **facebook/bart-large-mnli** (407 M parameters) – trained on the MNLI corpus of mostly formal English; (ii) **textattack/bert-base-MNLI** (110 M) – similarly fine-tuned on MNLI with little slang; (iii) **MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli** (418 M) – combines formal and informal data; and (iv) **roberta-large-mnli** (355 M) – pretrained on large formal corpora before MNLI tuning.

Model	Overall Accuracy
facebook/bart-large-mnli	0.6536
MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli	0.5939
textattack/bert-base-uncased-MNLI	0.4075
roberta-large-mnli	0.4487

Table 1: Stage 1 – Zero-shot slang detection via MNLI entailment.

4.2 Stage 2 – MNLI encoders with a task-specific head

Because the final layers of MNLI models are not dedicated classifiers, we add a two-way softmax head and train it on our slang dataset. The same four checkpoints are used:
(i) **facebook/bart-large-mnli**; (ii) **textattack/bert-base-uncased-MNLI**; (iii) **DeBERTa-v3-large-mnli**; and (iv) **roberta-large-mnli**.

Model	Overall Acc.	Slang Acc.	Non-Slang Acc.
facebook/bart-large-mnli	0.7348	0.6147	0.8154
DeBERTa-v3-large-mnli	0.7366	0.6124	0.8144
textattack/bert-base-MNLI	0.7219	0.5899	0.8098
roberta-large-mnli	0.6031	0.4005	0.7319

Table 2: Stage 2 – MNLI checkpoints with a fine-tuned classifier head.

4.3 Stage 3 – Chat models accessed via paid APIs

Budget constraints limited us to the lowest-cost commercial endpoints. We query each model with an identical system prompt and collect the YES/NO replies.

System prompt

```
You are an expert linguistic classifier.  For each user sentence, respond
with a single lowercase word:  'yes' if the sentence CONTAINS slang, 'no'
if it does NOT. Return ONLY 'yes' or 'no' .  no additional text.
```

Models evaluated are (i) **deepseek-llm-7b-chat** (7 B parameters) – cheapest commercial API, trained on a varied mix of formal and informal data; (ii) **meta-llama/Llama-2-7b-chat-hf** (7 B) – an open-source Meta chatbot fine-tuned for dialogue; and (iii) **gpt-4o** (parameter count undisclosed) – OpenAI's flagship model accessed through the Chat Completions API.

Model	Overall Acc.	Slang Acc.	Non-Slang Acc.
deepseek-llm-7b-chat	0.7034	0.7770	0.6500
meta-llama/Llama-2-7b-chat-hf	0.8494	0.7600	0.9900
gpt-4o	0.7379	0.4160	0.9470

Table 3: Stage 3 – Zero-shot slang detection with commercial chat models.

5 Explainability

5.1 TCAV test with Llama-2

To examine how large language models internally distinguish *slang* from *literal* usages of polysemous terms, we applied *Testing with Concept Activation Vectors* (TCAV; [5]) to the **Llama-2-7B** checkpoint. All analyses use mean-pooled hidden-state embeddings from the final transformer layer (31).

Concept construction. A linear probe was trained to separate embeddings of confirmed *fire-as-slang* occurrences from an equal number of literal instances. The probe's weight vector defines a one-dimensional *slang–literal* Concept Activation Vector (CAV). For every held-out sentence we report **(i)** the projection on this CAV, **(ii)** the model's predicted probability of the SLANG class, and **(iii)** the *directional derivative* $\partial y/\partial c$, obtained by back-propagating the SLANG probability (sigmoid output) with respect to the embedding and taking its dot-product with the CAV.

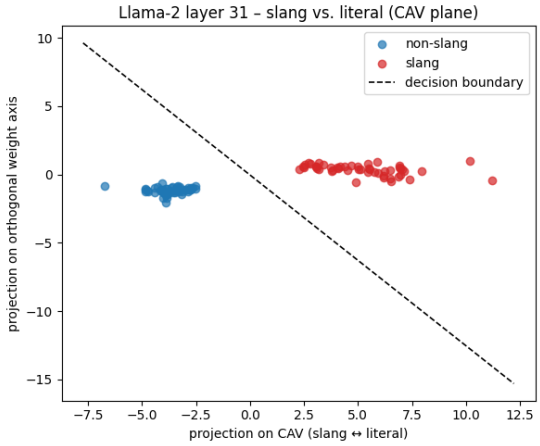


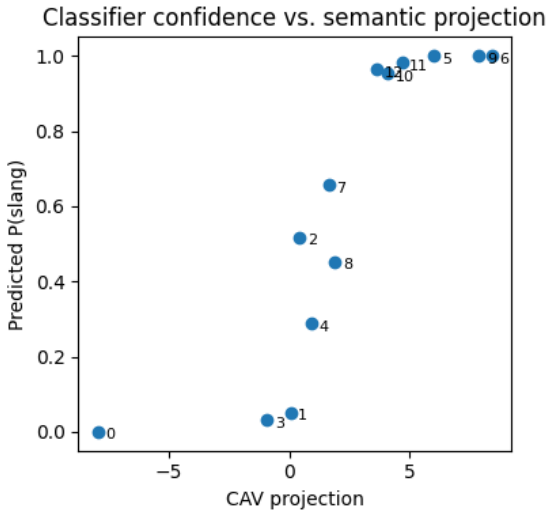
Figure 1: **CAV concept space visualization.** Sentence embeddings at Llama-2 layer 31 projected onto two axes: x-axis shows the learned slang ↔ literal CAV, y-axis shows the classifier weight component orthogonal to the CAV. Blue points represent literal usage, red points represent slang usage. The dashed line indicates the linear probe decision boundary. Clear cluster separation along the CAV demonstrates that the classifier acts essentially as a one-dimensional threshold on this semantic direction.

Results. Table 4 lists statistics for thirteen diverse *fire* sentences. The projections behave as expected: highly positive for clear slang ("This curry is fire", projection 7.88), strongly negative for literal fire (-7.95), and near-zero for ambiguous idioms ("Hold your fire", 0.44). The directional derivative is *always positive*, confirming that nudging the embedding further along the CAV consistently *increases* the model's slang logit. Crucially, its magnitude now varies by sentence: it peaks ($\approx$ 0.30) when the model is uncertain ($P(\text{slang}) \approx 0.5$) and shrinks to ≤ 0.002 when the model is already confident (probability ≈ 0 or 1). Thus TCAV not only identifies the concept direction but also quantifies *how much* each sentence's decision boundary depends on that concept.

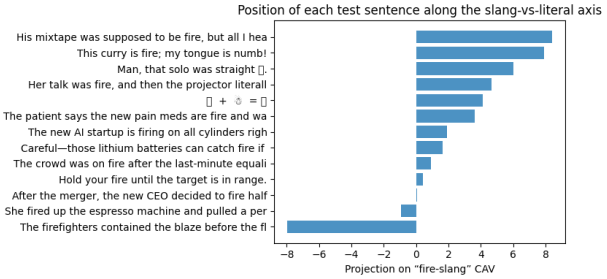
#	Sentence (truncated)	Pred.	$P(\text{slang})$	CAV proj.	$\partial y/\partial c$	Sign
0	The firefighters contained the blaze …	NO	0.000	−7.95	0.000002	↑
1	CEO decided to fire half the staff …	NO	0.051	0.05	0.058	↑
2	Hold your fire until the target …	SL	0.518	0.44	0.301	↑
3	She fired up the espresso machine …	NO	0.032	−0.94	0.037	↑
4	The crowd was on fire after the goal …	NO	0.287	0.93	0.247	↑
5	That solo was straight fire emoji	SL	0.999	6.02	0.002	↑
6	Mixtape was supposed to be fire, but …	SL	1.000	8.39	0.0004	↑
7	Lithium batteries can catch fire …	SL	0.657	1.63	0.272	↑
8	Startup is firing on all cylinders …	NO	0.453	1.89	0.299	↑
9	This curry is fire; my tongue is numb!	SL	1.000	7.88	0.0001	↑
10	fire emoji + snowman emoji = question mark emoji	SL	0.953	4.10	0.054	↑
11	Her talk was fire, then the projector …	SL	0.982	4.67	0.021	↑
12	Patient says the new pain meds are fire …	SL	0.963	3.62	0.043	↑

Table 4: TCAV statistics for *fire*. Projection and derivative values are rounded to three decimals.

Visualisations. Figure 2(a) shows the relationship between CAV projection and classifier confidence, with color intensity encoding directional derivative magnitude. The visualization confirms the monotonic relationship and highlights maximum sensitivity near the decision boundary. Figure 2(b) orders sentences along the slang-literal axis, clearly separating literal fire scenarios on the left from slang usages on the right.



(a) Classifier confidence vs. CAV projection

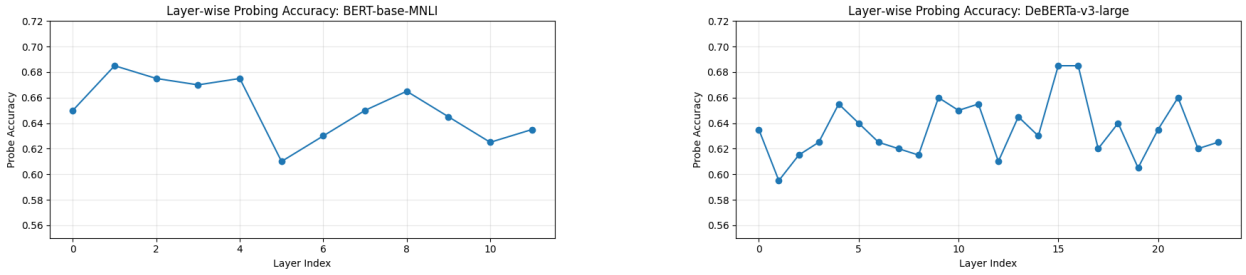


(b) Sentence ordering on slang-literal axis

Figure 2: **TCAV visualisations.** (a) Color gradient encodes directional derivative magnitude, highlighting maximum sensitivity near the decision boundary. (b) Sentences positioned along the learned concept axis show clear semantic separation.

5.2 Layers Slang Identifiers - Probing

After observing promising TCAV signals, we investigated *when* each model internalizes the notion of slang. Concretely, we trained a lightweight logistic regression classifier on the CLS token activations at each layer to predict "slang" vs. "non-slang" using a balanced dataset of 1,000 examples (500 slang and 500 non-slang sentences). We found that both models exhibit non-monotonic learning curves with distinct layer-wise patterns (Figure 3). BART-large-MNLI demonstrates peak slang discrimination ability in the middle layers (layers 5-8), achieving maximum accuracy of approximately 69.5%. In contrast, DeBERTa-v3-large shows more volatile behavior with multiple performance peaks, reaching its highest accuracy of approximately 68.5% in layers 15-16. Notably, both models exhibit performance degradation in their final layers, suggesting that the most discriminative representations for slang detection are not necessarily found in the uppermost layers. This indicates that middle-to-upper layers develop representations particularly well-suited for capturing linguistic style differences, while later layers may focus more on task-specific features from the models' MNLI training. Despite these different temporal dynamics in slang representation development, the overall difference in peak performance between the models is minimal ,indicating no meaningful correlation between *when* the model develops slang-sensitive representations and *how well* it ultimately performs on this linguistic style classification task.



(a) BERT-base-MNLI layer slang accuracy

(b) DeBERTa-v3-large layer slang accuracy

Figure 3: Per-layer slang classification accuracy for BERT and DeBERTa

5.3 Attention analysis with llama

Llama-2-7B-chat. After confirming that our TCAV probe reliably isolates a *slang* concept, we ran the same layer-wise attention study on **Llama-2-7B-chat**. For each sentence in a 1 k-example slice of OpenSubtitles, we found the sub-token containing the slang term, captured every self-attention map, and averaged attention *into* and *out of* that token across heads and sentences. Figure 4 shows a clear asymmetry: heads in the very first layer focus strongly on the slang token (some exceed 0.40 on our scale), while deeper layers flatten toward baseline. In contrast, attention *from* the slang token is widely dispersed-small pockets appear throughout the stack, with no single layer dominating. These patterns suggest that Llama detects the slang cue almost immediately, then redistributes that information rather than refining it in later layers.

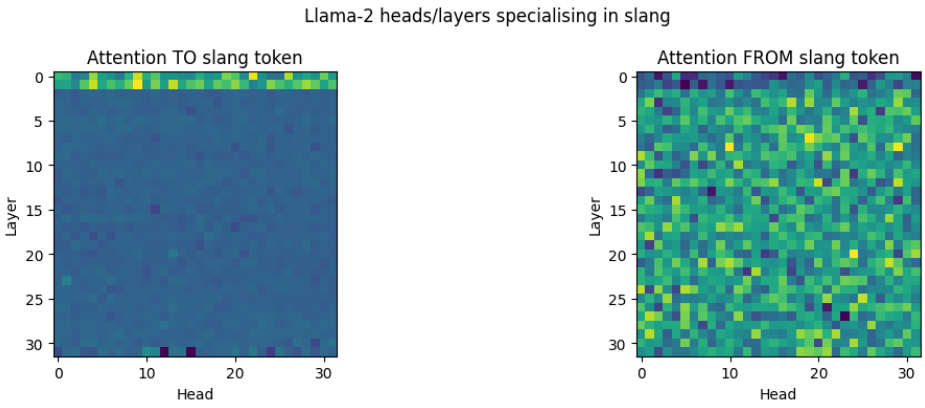


Figure 4: Layer–head attention specialization in **Llama-2-7B-chat**. Left: mean attention *into* slang tokens. Right: mean attention *from* slang tokens. Brighter cells indicate heads that allocate a larger share of attention to (or from) the slang term.

6 Discussion

Our findings reduce to three concise points: **(1) Behaviour.** With only a lightweight head, the two MNLI encoders reach 72–74% accuracy, while **Llama-2-7B-chat** achieves 85% straight out of the box. **(2) When the signal appears.** Layer-wise probing reveals non-monotonic learning curves with distinct patterns: BART-large-MNLI demonstrates peak slang discrimination in middle layers (5-8), while DeBERTa-v3-large shows volatile behavior with peaks in later layers (15-16). Both models exhibit performance degradation in final layers, suggesting optimal slang representations emerge in middle-to-upper layers rather than the final transformer blocks. **(3) Timing and performance.** Despite these different temporal dynamics in slang representation development, the overall difference in peak performance between models is minimal (under 1%), indicating no meaningful correlation between *when* the model develops slang-sensitive representations and *how well* it ultimately performs. Llama's higher score suggests that broader pre-training and model architecture, not the specific layer location of slang signals, drives performance gains.While our TCAV analysis was limited to Llama-2, the clear concept separation and meaningful directional derivatives suggest promise for this interpretability approach in understanding slang representations, warranting future multi-model validation.

In short, large LMs may develop peak slang discrimination early, late, or at multiple points in their processing stack-what matters for top-tier performance is the overall model capacity and training diversity, not the exact temporal emergence of slang-sensitive representations.

7 Conclusion and Future Work

We introduced a dual interpretability–behavioural framework for analysing slang ambiguity in LLMs. Our findings highlight the importance of vocabulary coverage and demonstrate the utility of TCAV for rapid audit. further research can explore:

- Enhancing the project with XAI-driven slang detection across various model architectures.
- Extending to multi-language slang by incorporating datasets in non-English languages.
- Generating synthetic slang data to evaluate models and reverse-engineer their internal representations.

- Investigating compositional slang phenomena and temporal drift in usage.
- Developing mitigation strategies, such as continual lexicon updating.

References

[1] Jacob Eisenstein, Amy O'Connor, and Noah A. Smith. Bad english in social media. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 359–369, Atlanta, GA, 2013.

[2] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*, pages 625–635, Florence, Italy, 2015. doi: 10.1145/2736277.2741627.

[3] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 345–363, Punta Cana, Dominican Republic, 2021. doi: 10.18653/v1/2021.emnlp-main.29.

[4] Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. Toward informal language processing: Knowledge of slang in large language models. *NAACL 2024 Long Papers*, 2024. to appear.

[5] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2668–2677, Stockholm, Sweden, 2018.