How Low Rank is Humor Recognition in LLMs?

Anonymous Author(s)

Abstract

Recent work in mechanistic interpretability has shown that large language models (LLMs) represent concepts like sentiment and truth as linear directions in their activation space. We investigate whether humor recognition shares this property by measuring the effective rank of humor-discriminating subspaces in GPT-2 (124M) and PYTHIA-410M (410M). Using PCA-constrained linear probes, mean difference directions, and LORA fine-tuning at varying ranks, we find a striking dissociation. When jokes are contrasted with stylistically dissimilar text (factual sentences), humor is essentially **rank-1**: a single linear direction achieves 99.8% classification accuracy, and LORA at rank 0 (a classification head alone) achieves 98.3%. However, when confounds are controlled by contrasting jokes with failed attempts at humor, linear probes achieve only $\sim$60% accuracy regardless of rank—barely above chance. This pattern holds across both models and stands in contrast to sentiment, which reaches 76.7% under similarly controlled conditions. Our results reveal that what LLMs linearly separate as "humor" is primarily text register and style, not humor understanding. True humor quality discrimination appears to require higher-rank or non-linear representations not easily extractable from frozen model activations, challenging the universality of the linear representation hypothesis for complex semantic properties.

1 Introduction

A growing body of work in mechanistic interpretability has revealed that large language models encode binary concepts—sentiment [Tigges et al., 2023], truth [Marks and Tegmark, 2023], and safety-relevant properties [Zou et al., 2023]—as linear directions in their activation space. These findings support the *linear representation hypothesis*: high-level concepts are represented as directions in the residual stream, recoverable by a single vector projection. A natural question is whether this linearity extends to more complex, subjective phenomena.

We examine humor. Humor is a compelling test case because it involves incongruity, surprise, and social context—properties that are arguably richer than positive-vs-negative sentiment. Prior work has shown that a single attention head in BERT specializes in humor detection [Peyrard et al., 2021] and that transformer-based classifiers achieve near-perfect accuracy on humor benchmarks [Weller and Seppi, 2019]. Yet no study has directly measured the *dimensionality* of humor recognition in LLM representations.

Our main question is: **how low-rank is humor recognition in LLMs?** Can humor be captured by a single direction (like sentiment), or does it require a higher-dimensional subspace?

We address this question through three complementary methods applied to GPT-2 (124M) and PYTHIA-410M (410M): (1) linear probes at PCA-constrained ranks from 1 to full dimensionality, (2) mean difference probes that project onto the rank-1 direction between class centroids, and (3) LORA fine-tuning at adapter ranks from 0 to 32. Critically, we evaluate these methods under two conditions: an *easy* setting that contrasts jokes with stylistically dissimilar factual text, and a *hard* setting that contrasts jokes with failed attempts at humor drawn from the same distribution.

Our results reveal a sharp dissociation. In the easy setting, humor is rank-1: a single direction achieves 99.8% accuracy in GPT-2 layer 8, and LoRA at rank 0 achieves 98.3%. In the hard setting, accuracy drops to $\sim$60% regardless of rank—barely above chance for binary classification. This gap shows that the "humor direction" discovered in the easy setting primarily captures text register (informal/conversational vs. formal/factual) rather than humor understanding. We compare against sentiment (SST-2), which achieves 76.7% accuracy under controlled conditions, confirming that humor quality is harder to linearly extract than even sentiment.

We make the following contributions:

- We provide the first systematic measurement of the effective rank of humor recognition in LLM representations, using three complementary methods across two models.
- We demonstrate that humor-style detection is rank-1 but that this result is driven by stylistic confounds, not humor comprehension.
- We show that humor quality discrimination (funny vs. unfunny jokes) is nearly linearly undetectable in frozen GPT-2 activations, achieving only $\sim$60% accuracy at any rank.
- We establish that LoRA at rank 0–1 suffices for humor-style classification but that this reflects surface features rather than genuine humor understanding.

2 Related Work

Linear representations in LLMs. The linear representation hypothesis posits that high-level concepts are encoded as directions in a model's residual stream. Tigges et al. [2023] demonstrated that sentiment is linearly represented in GPT-2 and the Pythia family, with five independent methods (mean difference, K-means, logistic regression, PCA, and DAS) converging to the same direction (cosine similarity 79–99%). Ablating this direction removes 76% of sentiment classification accuracy. Marks and Tegmark [2023] extended this finding to truth, showing that LLMs form a linear "truth direction" that generalizes across diverse factual statement types. Zou et al. [2023] introduced Representation Engineering, identifying linear directions for honesty, fairness, and harmlessness. More recently, Engels et al. [2024] challenged the universality of this hypothesis by demonstrating that some features (e.g., periodic concepts like days of the week) are represented as multi-dimensional, non-linear structures. Our work extends this line of inquiry to humor, testing whether it behaves as a simple linear feature or a more complex representation.

Computational humor recognition. Transformer-based humor detection has achieved near-perfect accuracy on standard benchmarks. Weller and Seppi [2019] reported 98.6% accuracy on the SHORT JOKES dataset using a fine-tuned BERT model. Meaney et al. [2021] introduced the HaHackathon benchmark (SemEval-2021 Task 7), where top systems reached F1 scores of 0.97 for binary humor detection. However, Inácio et al. [2023] demonstrated that these classifiers rely heavily on stylistic features (punctuation, question words) rather than deep humor understanding—a finding our controlled experiments directly confirm at the representation level. Xie et al. [2020] modeled humor through incongruity theory using GPT-2 perplexity, suggesting that humor detection may partially reduce to distributional statistics.

Mechanistic analysis of humor in transformers. Peyrard et al. [2021] provided the first mechanistic analysis of humor processing in transformers, discovering a single "laughing head" (attention head 10-6 in BERT) that specializes in attending to humor-bearing tokens. This finding—that humor concentrates in a single model component—is suggestive of low-rank structure, though Peyrard et al. analyzed attention patterns rather than representation dimensionality. Li et al. [2022] showed that prompting achieves comparable performance to fine-tuning for humor recognition, implying that pre-trained models already contain humor-relevant information. Our work complements these studies by directly measuring the dimensionality of humor in hidden representations rather than in attention patterns or output behavior.

Low-rank structure in fine-tuning. Aghajanyan et al. [2020] showed that NLP fine-tuning tasks have remarkably low intrinsic dimensionality: RoBERTa-Large reaches 90% of full performance on MRPC with only 200 trainable parameters. Pre-training monotonically decreases intrinsic dimension, and larger models have lower intrinsic dimension. Hu et al. [2021] exploited this insight to develop LoRA, which decomposes weight updates as $\Delta \boldsymbol{W} = \boldsymbol{B}\boldsymbol{A}$ where $\boldsymbol{B} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{A} \in \mathbb{R}^{r\times k}$ with $r \ll \min(d, k)$. Wu et al. [2024] pushed this further with LoReFT, achieving competitive results with 10–50$\times$ fewer parameters than LoRA by operating directly in representation space. We

Table 1: Dataset conditions. All conditions are class-balanced. Reddit jokes are filtered to remove `[deleted]` posts.

Condition	Positive	Negative	Source	N_{train} / N_{test}
Easy	Short jokes	Factual sentences	Kaggle + manual	1,200 / 400
Hard-1	Short jokes	Low-score Reddit	HuggingFace	1,000 / 300
Hard-2	High-score Reddit	Low-score Reddit	HuggingFace	1,000 / 300
Sentiment	SST-2 positive	SST-2 negative	HuggingFace	1,000 / 300

use LoRA rank as an independent measure of humor recognition dimensionality, complementing our probing-based measurements.

Probing methodology. Hewitt and Liang [2019] introduced control tasks to validate probing results, ensuring that probe accuracy reflects genuine linguistic knowledge rather than memorization capacity. Burns et al. [2022] developed Contrast Consistent Search (CCS) for finding latent knowledge directions without supervision. Our probing methodology follows Tigges et al. [2023] but extends it with rank-constrained evaluation and controlled confound experiments.

3 Methodology

We measure the effective rank of humor recognition through three complementary methods: rank-constrained linear probing, mean difference probing, and LoRA fine-tuning at varying adapter ranks. We apply each method under controlled experimental conditions designed to disentangle humor detection from text-style detection.

3.1 Models

We use two autoregressive language models:

- GPT-2 [Radford et al., 2019]: 124M parameters, 12 layers, hidden dimension 768. Our primary model, chosen for its extensive use in mechanistic interpretability research [Tigges et al., 2023, Marks and Tegmark, 2023, Nanda et al., 2023].
- PYTHIA-410M [Biderman et al., 2023]: 410M parameters, 24 layers, hidden dimension 1024. Used to test whether humor rank changes with model scale.

3.2 Datasets

We construct five experimental conditions spanning two difficulty levels (table 1).

Easy task (humor style). We contrast short jokes from the Kaggle SHORT JOKES dataset with hand-crafted factual sentences (e.g., "The speed of light is approximately 299,792 kilometers per second"). Jokes are filtered to 10–200 characters to reduce trivial length cues. This condition tests whether the model linearly separates joke-style text from non-joke text.

Hard tasks (humor quality). We design two conditions that control for text register. In **Hard-1**, we contrast SHORT JOKES with low-scoring Reddit jokes (score ≤ 2)—both are attempts at humor, differing only in quality. In **Hard-2**, we contrast high-scoring Reddit jokes (score ≥ 50) with low-scoring ones. These conditions test whether the model represents humor *quality* rather than humor *style*.

Sentiment baseline. We probe SST-2 sentiment (positive vs. negative movie reviews) as a calibration benchmark, since sentiment is known to be rank-1 [Tigges et al., 2023].

All datasets are balanced 50/50 between classes, filtered by character length (20–200), and split with a fixed seed of 42 for reproducibility.

3.3 Activation Extraction

For each input text, we perform a forward pass through the model and extract the hidden state $\boldsymbol{h}_\ell \in \mathbb{R}^d$ at the last non-padding token position for every layer $\ell \in \{0, 1, \ldots, L\}$, where L is the number of transformer layers and layer 0 denotes the embedding layer output. We collect activations for all training and test examples, yielding activation matrices $\boldsymbol{H}_\ell^+ \in \mathbb{R}^{n \times d}$ (positive class) and $\boldsymbol{H}_\ell^- \in \mathbb{R}^{n \times d}$ (negative class) at each layer.

3.4 Rank-Constrained Linear Probing

To measure classification accuracy as a function of representation rank, we apply PCA dimensionality reduction before logistic regression. For a given rank k and layer ℓ:

1. Standardize activations (zero mean, unit variance per feature).
2. Fit PCA with k components on training activations.
3. Transform both training and test activations to k dimensions.
4. Train logistic regression on the k-dimensional training set.
5. Evaluate accuracy on the k-dimensional test set.

We sweep $k \in \{1, 2, 4, 8, 16, 32, 64, d\}$ and $\ell \in \{0, \ldots, L\}$, yielding an accuracy surface over rank and layer. If humor is rank-r, accuracy should saturate at $k = r$.

3.5 Mean Difference Probe

Following Tigges et al. [2023], we compute the rank-1 humor direction as the unit vector between class centroids:

$$\boldsymbol{d}_\ell = \frac{\bar{\boldsymbol{h}}_\ell^+ - \bar{\boldsymbol{h}}_\ell^-}{\|\bar{\boldsymbol{h}}_\ell^+ - \bar{\boldsymbol{h}}_\ell^-\|} \tag{1}$$

where $\bar{\boldsymbol{h}}_\ell^+$ and $\bar{\boldsymbol{h}}_\ell^-$ are the mean activations for the positive and negative classes at layer ℓ. We classify test examples by projecting onto $\boldsymbol{d}_\ell$ and thresholding at the median projection value. This provides a direct rank-1 accuracy measure that is independent of probe capacity.

3.6 LoRA Fine-Tuning

We fine-tune GPT-2 with LoRA adapters [Hu et al., 2021] injected into the attention projection matrices (`c_attn`, `c_proj`) and add a binary classification head on top of the last token's hidden state. We sweep adapter rank $r \in \{0, 1, 2, 4, 8, 16, 32\}$, where $r = 0$ corresponds to training only the classification head (1,536 parameters) with the base model frozen. All experiments use learning rate 2×10^{-4}, batch size 32, and 5 training epochs.

The LoRA rank at which accuracy saturates provides a complementary measure of humor recognition dimensionality: if the weight updates needed for humor detection are rank-r, then LoRA at rank r should suffice.

3.7 Evaluation

We report classification accuracy on held-out test sets. All experiments use a fixed random seed (42) for reproducibility. We identify the *best layer* for each condition (the layer maximizing test accuracy at the best rank) and report accuracy across ranks at that layer, as well as the best rank-1 accuracy across all layers.

4 Results

4.1 Easy Task: Humor Style is Rank-1

Figure 1 summarizes our main probing results on the easy task (jokes vs. factual text) using GPT-2.

Rank-1 probes achieve near-perfect accuracy. The mean difference probe achieves 99.8% accuracy at layer 8—a single linear direction separates jokes from factual sentences almost perfectly

Table 2: Easy task (jokes vs. factual): rank-1 probe accuracy by layer for GPT-2. The mean difference and PCA rank-1 probes produce nearly identical results. Bold indicates the best layer.

Layer	Mean Diff	Rank-1	Best Rank	Best Acc
0	0.964	0.964	1	0.964
1	0.974	0.980	32	0.996
2	0.990	0.988	32	0.996
3	0.983	0.988	64	0.999
4	0.985	0.989	32	0.999
5	0.989	0.989	16	0.999
6	0.990	0.989	16	0.999
7	0.995	0.996	8	1.000
8	**0.998**	**0.998**	**8**	**1.000**
9	0.994	0.995	32	1.000
10	0.994	0.995	32	1.000
11	0.989	0.995	32	1.000
12	0.661	0.981	8	0.998

Table 3: Easy task: probe accuracy at varying ranks for GPT-2 layer 8. Accuracy saturates at rank 1, confirming that humor-style information is one-dimensional.

Rank	Accuracy
1	0.998
2	0.996
4	0.996
8	0.998
16	0.999
32	1.000
64	0.999
768	0.999

(table 2). Accuracy is high even at the embedding layer (96.4%) and peaks in the middle layers (7–9), consistent with the finding that semantic features crystallize at intermediate layers [Tigges et al., 2023].

Higher rank provides negligible improvement. At the best layer (layer 8), increasing the probe rank from 1 to 768 (full dimensionality) improves accuracy from 99.8% to 99.9% (table 3). The accuracy curve is essentially flat beyond rank 1, confirming that humor-style information concentrates in a single dimension.

4.2 LoRA Fine-Tuning Confirms Low Rank

LoRA fine-tuning on the easy task corroborates the probing results (table 4). Training only the classification head ($r = 0$, 1,536 parameters) achieves 98.3% accuracy—the frozen GPT-2 representations already encode humor-style information that a simple linear layer can exploit. Adding

Figure 1: Humor Recognition Rank Analysis (GPT-2 Small, Jokes vs. Factual Text)

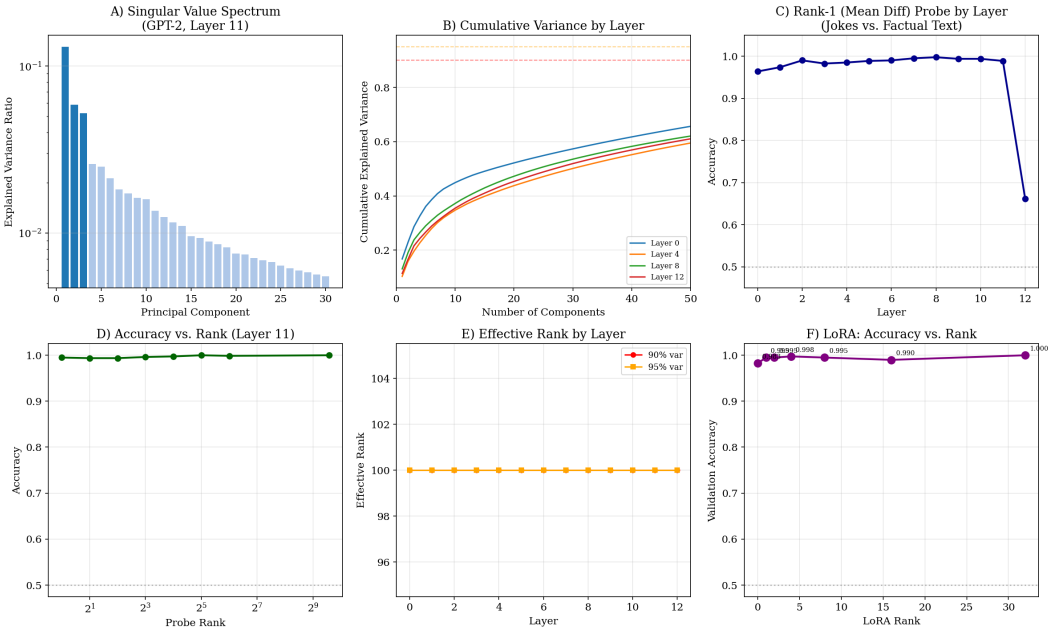

Figure 1: Main results for the easy task (jokes vs. factual text) using GPT-2. *(Left)*PCA eigenvalue spectrum showing rapid variance decay. *(Right)*Probing accuracy across layers and ranks, confirming rank-1 saturation at intermediate layers.

Table 4: LoRA fine-tuning results on the easy task with GPT-2. Rank 0 denotes training only the classification head. Accuracy saturates at rank 1–4.

LoRA Rank	Trainable Params	Val Accuracy
0	1,536	0.983
1	102,912	0.995
2	204,288	0.995
4	407,040	**0.998**
8	812,544	0.995
16	1,623,552	0.990
32	3,245,568	**1.000**

a rank-1 LoRA adapter (102,912 parameters) increases accuracy to 99.5%, and performance saturates at rank 4 (99.8%). The highest rank tested ($r = 32$) achieves 100% but uses 3.2M trainable parameters—a 2,000× increase for a 1.7 percentage point gain.

4.3 Hard Tasks: Humor Quality is Not Low-Rank

Controlling for style collapses performance. When we replace factual sentences with low-scoring Reddit jokes as the negative class (Hard-1), the best rank-1 probe accuracy drops from 99.8% to 59.7% (table 5). When both classes are Reddit jokes differing only in score (Hard-2), the best rank-1 accuracy is 56.7%. These results are barely above the 50% chance baseline for balanced binary classification.

Full-rank probes offer no rescue. Increasing probe rank to full dimensionality (768) does not help: the best accuracy across all ranks and layers reaches only 60.0% for Hard-1 and 61.0% for Hard-2. The humor-quality signal, if present in the representations, is not linearly accessible at any rank.

Sentiment comparison. As a calibration, we probe SST-2 sentiment under the same protocol. The best rank-1 accuracy is 64.0% (layer 11), and the best full-rank accuracy is 76.7% (layer 11).

Figure 5: LoRA Fine-tuning for Humor Detection (GPT-2, Jokes vs. Factual)

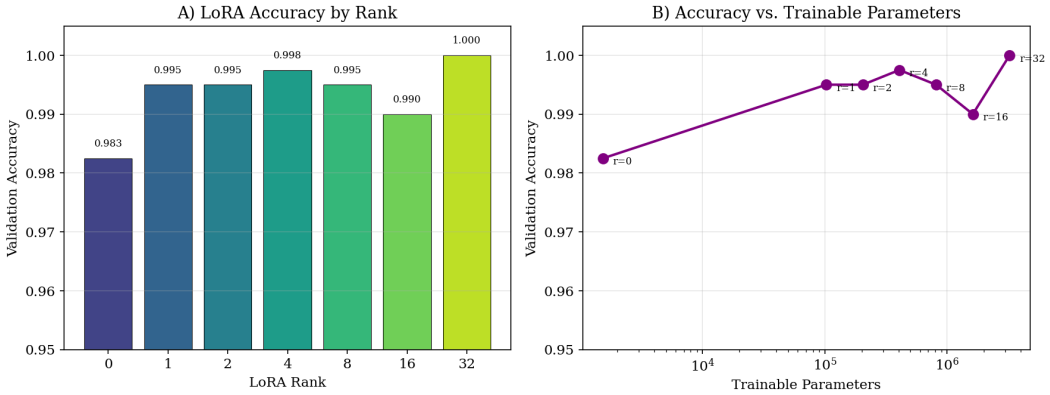

Figure 2: Detailed LoRA rank analysis. *(Left)*Accuracy vs. LoRA rank showing early saturation. *(Right)*Trainable parameter count grows linearly with rank, but accuracy gains diminish rapidly after rank 1.

Table 5: Hard task and sentiment results. Controlling for stylistic confounds collapses humor detection accuracy to near-chance. Sentiment remains partially linearly separable. Best results per task in bold.

Task	Best Rank-1 Acc	Layer	Best Full-Rank Acc	Layer
Easy (jokes vs. factual)	**0.998**	8	**1.000**	7
Hard-1 (jokes vs. Reddit)	0.597	7	0.600	10
Hard-2 (high vs. low Reddit)	0.567	9	0.610	12
Sentiment (SST-2)	0.640	11	0.767	11

Sentiment is more linearly separable than humor quality but less separable than humor style—and it requires higher rank, consistent with prior work showing sentiment needs multiple dimensions for full recovery [Tigges et al., 2023].

4.4 Cross-Model Validation: Pythia-410M

We replicate the easy-task experiments on PYTHIA-410M to test whether the rank-1 pattern generalizes across architectures and scales (table 6). The results are consistent: the mean difference probe achieves 100% accuracy at layer 20 (of 24), and accuracy builds steadily through the layers. As with GPT-2, rank-1 is sufficient—higher ranks provide at most marginal improvement. The larger model achieves perfect separation at a proportionally similar depth ($\sim$83% of layers for PYTHIA-410M vs. $\sim$67% for GPT-2), suggesting that the humor-style direction is a robust feature of autoregressive language models, not an artifact of model size or architecture.

5 Discussion

5.1 The Humor Direction is a Style Direction

Our central finding is that the rank-1 "humor direction" in GPT-2 and PYTHIA-410M primarily captures text register—the difference between informal, conversational joke-style text and formal, factual text—rather than humor understanding. Three pieces of evidence support this interpretation.

First, the 99.8% rank-1 accuracy on the easy task is *suspiciously high*. Even sentiment, a simpler binary concept with well-established linear structure [Tigges et al., 2023], does not achieve this level of performance from frozen representations under our probing protocol (76.7% at best). The near-perfect accuracy suggests the probe exploits a surface-level signal rather than a semantic one.

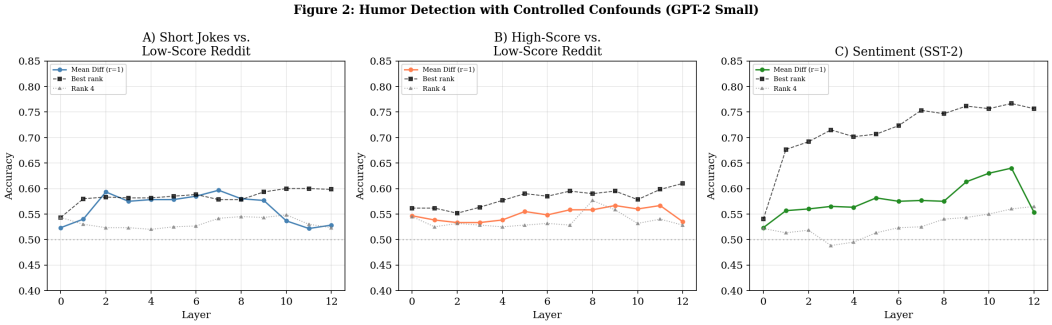

Figure 3: Hard task results. Probe accuracy across layers for the controlled conditions. Unlike the easy task, no layer or rank achieves substantially above-chance accuracy for humor quality discrimination.

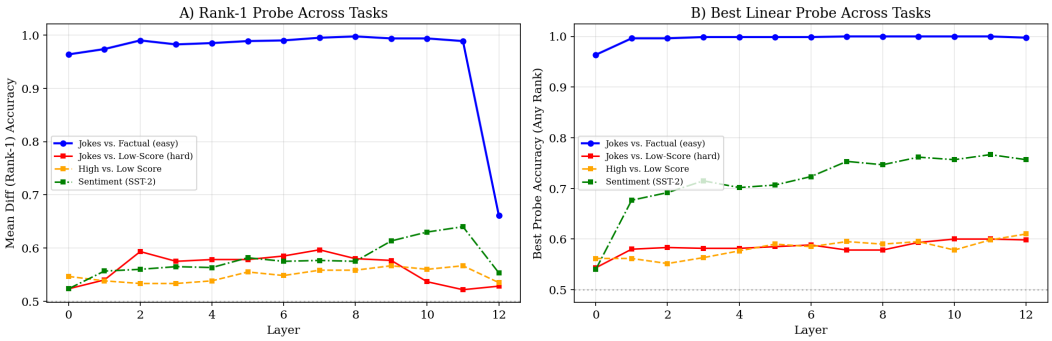

Figure 4: Comparison across conditions. *(Left)*Easy vs. hard tasks showing the dramatic accuracy drop when stylistic confounds are controlled. *(Right)*Humor quality vs. sentiment, showing that humor quality is less linearly separable than sentiment.

Second, accuracy drops to near-chance (∼60%) when both classes share the same text register. In Hard-1 (jokes vs. unfunny Reddit jokes) and Hard-2 (popular vs. unpopular Reddit jokes), the stylistic confound is removed, and the probe fails. If the "humor direction" captured genuine humor recognition, it should maintain discriminative power when applied to texts that differ in humor quality but not in style.

Third, LoRA at rank 0—training only a 1,536-parameter classification head on frozen activations—achieves 98.3% on the easy task. This confirms that the relevant information is already present in the frozen representations and does not require learning new features. The information is so accessible that a linear layer suffices, consistent with a text-register signal that is strongly represented in the base model.

5.2 Implications for the Linear Representation Hypothesis

Our results add nuance to the linear representation hypothesis. The hypothesis holds for *stylistic* properties: text register is linearly and low-rank represented. However, it does not straightforwardly extend to *humor quality*, a more complex semantic property. This aligns with Engels et al. [2024], who showed that not all features in LLMs are linear. Humor quality may require non-linear representations, multi-dimensional manifolds, or features that are not well-captured by single-token activations.

The distinction between "linearly detectable" and "genuinely understood" is critical for interpretability research. A high-accuracy linear probe does not guarantee that the model encodes the target concept; it may instead reflect a confound that correlates with the concept in the training data [Hewitt and Liang, 2019]. Our controlled experiments provide a template for testing this distinction: measure probing accuracy both with and without stylistic confounds.

Table 6: Cross-model validation: easy task results for PYTHIA-410M. The rank-1 pattern replicates, with perfect accuracy at layer 20.

Layer	Mean Diff Acc	Best Rank	Best Acc
0	0.759	2	0.759
5	0.971	32	0.998
10	0.994	4	0.999
15	0.998	8	0.999
20	**1.000**	**4**	**1.000**
24	0.990	16	0.999

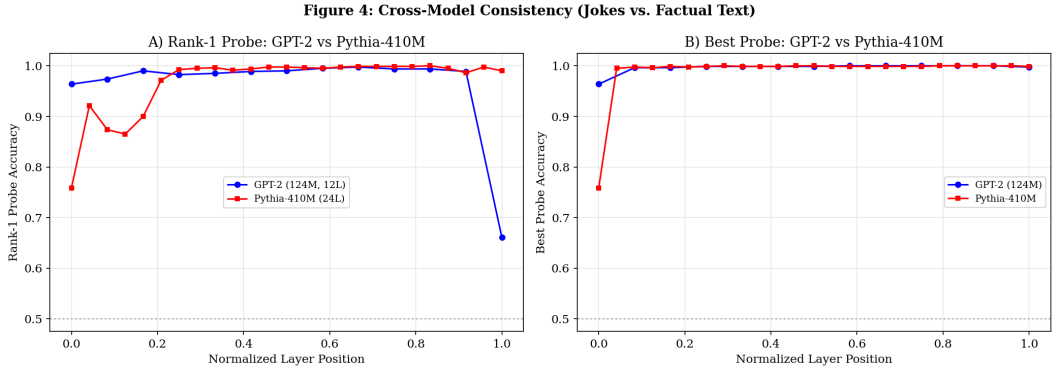

Figure 5: Cross-model comparison. GPT-2 (left) and PYTHIA-410M (right) show the same pattern on the easy task: rank-1 accuracy peaks in middle-to-late layers, confirming that humor-style detection is a robust, low-rank feature.

5.3 Implications for Practical Applications

For practitioners building humor classifiers, our results have a clear message: LORA at rank 1 is sufficient to distinguish jokes from non-jokes (99.5% accuracy), but this classifier detects joke *style*, not joke *quality*. Building a system that distinguishes genuinely funny from unfunny text likely requires either (1) larger models with richer representations, (2) non-linear probes or classifiers, or (3) fine-tuning that creates new humor-relevant features rather than exploiting existing ones.

5.4 Limitations

Non-humor text quality. Our factual sentences are hand-crafted and maximally dissimilar from jokes. A more naturalistic non-humor set (e.g., casual non-joke conversation) might yield intermediate results between the easy and hard conditions.

Reddit score as humor proxy. Reddit upvotes reflect many factors beyond humor quality—timing, subreddit norms, controversy, and audience demographics. Low-scoring jokes may not be unfunny; they may simply be unseen. A cleaner humor-quality signal might come from controlled human ratings, such as the Unfun.me dataset [Horvitz et al., 2024, Peyrard et al., 2021].

Model scale. GPT-2 (124M) and PYTHIA-410M (410M) are relatively small by modern standards. Larger models (7B+) may develop richer humor representations that are more linearly accessible. Our negative result for humor quality should be interpreted as applying to these specific model sizes.

Activation position. We extract activations at the last token position only. Humor information may be distributed across multiple positions, particularly in setup-punchline structures where the humor resides in the relationship between distant tokens.

Binary framing. We test binary humor detection (funny vs. not funny). Humor is graded and multi-dimensional—puns, absurdity, irony, and dark humor may occupy distinct subspaces. A finer-grained analysis might reveal low-rank structure within humor subtypes even if the overall humor space is high-rank.

Language and culture. All datasets are English-language. Humor is highly culture- and language-dependent, and our findings may not generalize to other linguistic contexts.

6 Conclusion

We measured the effective rank of humor recognition in LLM representations using rank-constrained linear probes, mean difference directions, and LoRA fine-tuning across GPT-2 and Pythia-410M. Our results reveal that the answer depends critically on what "humor recognition" means.

Distinguishing joke-style text from non-joke text is **rank-1**: a single linear direction achieves 99.8% accuracy, and LoRA at rank 0 achieves 98.3%. However, this direction captures text register, not humor understanding. When stylistic confounds are controlled—contrasting jokes with failed attempts at humor—linear probes achieve only $\sim$60% accuracy regardless of rank, showing that humor quality is nearly linearly undetectable in frozen model representations.

These findings have three implications. For **interpretability**, they caution that high-accuracy linear probes may reflect stylistic confounds rather than genuine concept encoding; controlled experiments are essential for validating linear representation claims. For **applications**, they show that LoRA at rank 1 suffices for joke-style detection, but building a humor quality classifier requires more expressive methods. For **humor research**, they suggest that LLM humor benchmarks may substantially overestimate genuine humor understanding due to uncontrolled stylistic differences between humor and non-humor text.

Future work. Promising directions include testing with larger models (7B+) to determine whether humor quality becomes linearly separable at scale, using non-linear probes to search for low-dimensional but curved humor manifolds, and applying sparse autoencoders [Elhage et al., 2022] to identify humor-related features in a more fine-grained manner. The Unfun.me paired dataset [Peyrard et al., 2021, Horvitz et al., 2024]—which provides minimal pairs of funny and unfunny headlines—offers an ideal test bed for eliminating stylistic confounds entirely.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Joshua Engels, Isaac Liao, and Max Tegmark. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

Zachary Horvitz, Jingru Chen, Rahul Aditya, Kanishk Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. Getting serious about humor: Crafting humor datasets with unfunny LLMs. *arXiv preprint arXiv:2403.00794*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Marcio Lima Inácio, Gabriela Wick-Pedro, and Hugo Gonçalo Oliveira. What do humor classifiers learn? an attempt to explain transformer-based humor recognition models. *Information*, 14(5): 300, 2023.

Junze Li, Mengjie Zhao, Yubo Xie, Antonis Maronikolakis, Pearl Pu, and Hinrich Schütze. This joke is [MASK]: Recognizing humor and offense with prompting. *arXiv preprint arXiv:2209.12118*, 2022.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

J A Meaney, Steven R Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. SemEval-2021 task 7: HaHackathon, detecting and rating humor and offense. *arXiv preprint arXiv:2105.13602*, 2021.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Maxime Peyrard, Beatriz Borges, Kristina Gligorić, and Robert West. Laughing heads: Can transformers detect what makes a sentence funny? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Curt Tigges, Oskar John Lincoln Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

Orion Weller and Kevin Seppi. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. ReFT: Representation finetuning for language models. *arXiv preprint arXiv:2402.14700*, 2024.

Yubo Xie, Junze Li, and Pearl Pu. Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition. *arXiv preprint arXiv:2012.12007*, 2020.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyber, Dawn Song, Matt Fredrikson, J Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.