MIXTURE-OF-EXPERTS ROUTING RESPONSES ARE STRONG SIGNALS FOR AI-GENERATED TEXT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalizing AI-Generated Text Detection to unseen domains, generators, and generation tasks remains challenging. We introduce a simple yet effective detector that uses the native routing responses of a frozen Mixture-of-Experts (MoE) language model. Although pretrained without any AI-Generated Text Detection objective, these routers produce responses that could be used for distinguishing authorship. Our method averages router logits over tokens, concatenates the resulting features across layers, and trains a single linear classifier while keeping the observer frozen. With GPT-OSS-20B, the classifier contains only 769 trainable parameters and is trained on 500 human-AI text pairs. On MIRAGE, it achieves AUROCs of 0.9663, 0.9909, and 0.9937 in cross-domain, cross-generator, and cross-task evaluations, respectively, outperforming all evaluated baselines. Controlled analyses show that detection performance is largely preserved when removing logit shifts shared across experts, but declines when expert identity is removed. Even discrete expert-selection frequencies retain a strong detection signal. Evaluations across nine frozen MoE observers further show that this routing footprint recurs across model families. These findings show that native MoE routing exposes an authorship footprint carried by relative, expert-specific preferences, and that a minimal linear readout of routing responses supports strong detection across diverse observers.

1 INTRODUCTION

As Large Language Models (LLMs) continue to advance (Achiam et al., 2023; Grattafiori et al., 2024; Liu et al., 2024; Yang et al., 2025; Zeng et al., 2026; Xiao et al., 2026), AI-generated text is becoming increasingly difficult to distinguish from human writing (Sadasivan et al., 2023; Weber-Wulff et al., 2023; Li et al., 2024), particularly when detectors are applied to unseen generators, domains, or generation settings (Dugan et al., 2024; Wu et al., 2024; Fu et al., 2025). A central challenge is therefore to identify detection signals that generalize reliably across such distribution shifts.

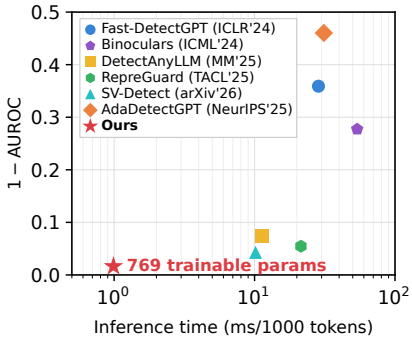


Figure 1: Mean detection performance versus runtime; lower left is better.

A prominent class of AI-text detectors asks how a language model *scores* an observed text (Gehrmann et al., 2019; Mitchell et al., 2023; Su et al., 2023). For instance, methods such as Fast-DetectGPT (Bao et al., 2024), Binoculars (Hans et al., 2024), and AdaDetectGPT (Zhou et al., 2026) construct detection signals from likelihoods, probability statistics, or differences in scoring behavior across language models.

A more recent line of work turns to the internal representations of pretrained language models (Yu et al., 2024; Liang & Li, 2026). Methods such as RepreGuard (Chen et al., 2025b) and SV-Detect (Vishnyakov & Gaintseva, 2026) show that hidden states contain discriminative information about whether a text is human- or AI-generated. These findings suggest that authorship information is already embedded in representations acquired during language-model pretraining. Yet extract-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ing such information from high-dimensional, general-purpose representations typically requires detection-specific projections or readout mechanisms.

Mixture-of-Experts (MoE) models (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Jiang et al., 2024) provide a natural setting for reading authorship information through a pretrained mechanism. At each MoE layer, the router takes the contextual hidden state of a token and projects it into a compact set of expert preferences, which directly determines how computation is allocated (Jiang et al., 2024; Dai et al., 2024). Importantly, the router is learned jointly with the model under the language-modeling objective, rather than being introduced specifically for AI-generated text detection (Fedus et al., 2022; Liu et al., 2024).

Unlike raw hidden states, routing responses lie in a structured, low-dimensional space with explicit computational meaning (Ye et al., 2026). They describe which experts the model prefers for each contextual token, making routing a natural observation window into how the pretrained model internally responds to a text (Xue et al., 2024; Muennighoff et al., 2025; Arnold et al., 2024). We therefore ask: ***can authorship information already be exposed through these native routing responses strongly enough to be recovered with only a minimal readout?***

Importantly, we do not view routing as a new independent source of authorship information. Because routing responses are computed directly from contextual hidden states, any authorship signal they expose must ultimately originate from representations learned during pretraining.

There is no a priori reason for this signal to remain readily accessible after routing. The router is only a low-dimensional projection of a much richer hidden representation, and it is optimized for expert allocation rather than for preserving directions useful for authorship discrimination. Thus, even if authorship is strongly represented in hidden states, it may be attenuated, entangled, or discarded by the routing bottleneck. We study routing under an intentionally restrictive readout, in order to minimize the possibility that detection performance arises from task-specific representation learning in the readout, rather than from information already exposed by routing (Alain & Bengio, 2017; Hewitt & Liang, 2019; Belinkov, 2022). Specifically, we freeze the MoE observer, average its routing responses across tokens, and fit only a single linear classifier on top. Under this constraint, strong detection performance would provide direct evidence that authorship information is already readily accessible in routing space, rather than being constructed by the detection-specific readout itself. We refer to the resulting authorship-dependent pattern in routing responses as a *routing footprint*. Surprisingly, we find that this routing footprint is remarkably strong.

With GPT-OSS-20B (Agarwal et al., 2025) as the frozen observer, the entire learned component contains only **769 parameters** and is fitted using merely **500 paired human/AI examples**. Nevertheless, on MIRAGE (Fu et al., 2025), mean router logits achieve AUROCs of **0.9663**, **0.9909**, and **0.9937** under our cross-domain, cross-generator, and cross-task generalization evaluation settings, respectively. These results provide strong evidence that the pretrained routing projection preserves authorship-related structure from the underlying hidden representations in a form that is readily accessible to a linear readout.

To understand *why authorship remains so readily readable after the pretrained routing projection*, we examine which aspects of routing carry authorship information, its variation across depth and context, whether richer cross-layer or temporal structure contributes beyond mean expert preferences, and its recurrence across MoE observers.

Our contributions are summarized as follows:

1. **We uncover an authorship routing footprint in pretrained MoE models.** Human and AI-generated texts elicit linearly distinguishable routing responses, even though the router is trained solely for language modeling without authorship supervision.

2. **We demonstrate strong detection generalization with a minimal routing readout.** Token-mean router logits achieve the highest AUROC among the evaluated baselines in all three settings using a frozen observer and a single linear classifier.

3. **We characterize the structure and generality of the routing footprint.** Controlled probes reveal a strong signal in relative expert preferences and discrete selections, its variation across depth and context, and strong linear readability across nine frozen MoE observers.

2 Related Work

2.1 AI-Generated Text Detection

Statistical detectors identify generation artifacts through language-model scoring, including probability curvature and comparative scoring (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024). Recent work extends this approach through probability estimation and implicit reward signals (Bao et al., 2025; Liu et al., 2026). Learning-based methods improve detection through task-specific objectives, stylistic adaptation, and learned statistical criteria (Fu et al., 2025; Chen et al., 2025a; Zhou et al., 2026; Verma et al., 2024; Hu et al., 2023). DEER learns instance-adaptive expert routing using detection rewards within a task-specific MoE detector (Ma et al., 2025). Benchmarks nevertheless highlight persistent challenges under shifts in domains, generators, and editing operations (Dugan et al., 2024; Wu et al., 2024; Fu et al., 2025).

Representation-based detectors exploit writing-style embeddings and intermediate activations (Rivera Soto et al., 2024; Yu et al., 2024). Directional scores, supervised subspaces, and representation steering further expose authorship-related differences in these representations (Chen et al., 2025b; Vishnyakov & Gaintseva, 2026; Xiong et al., 2026; Liang & Li, 2026). These findings motivate examining whether such information survives projections already learned during pretraining. We investigate this question through native MoE routing responses, using a frozen observer, token averaging, and a linear readout for cross-distribution detection.

2.2 Interpreting LLM Activations and MoE Routing

Interpretability research examines the information encoded in language-model activations. Concept directions connect linear representations to probing and steering (Park et al., 2023; Zou et al., 2023). Sparse autoencoders decompose activations into features amenable to scalable interpretation (Gao et al., 2025; Paulo et al., 2025; Templeton et al., 2026). Recent evaluations emphasize distinguishing feature interpretability from reconstruction quality and downstream predictive utility (Karvonen et al., 2025; Kantamneni et al., 2025).

MoE analyses reveal structured token–expert associations and expert specialization in pretrained routing (Xue et al., 2024; Muennighoff et al., 2025; Ye et al., 2026). Routing also reflects linguistic categories and contextual information (Antoine et al., 2025; Arnold et al., 2024). Studies of shared expert coalitions challenge simple correspondences between experts and semantic domains (Wang et al., 2026). We extend these analyses to authorship, probing signal retention in expert preferences and discrete selections, its variation across depth and context, and its recurrence across observers.

3 Methodology

3.1 Native Routing Responses As a Pretrained Projection

Let $x = \{x_1, x_2, \ldots, x_S\}$ denote a text consists S tokens, and $y \in \{0, 1\}$ denote its authorship label, where $y = 0$ corresponds to human-written text and $y = 1$ corresponds to AI-generated text.

We use a pretrained causal MoE language model M_θ as an *observer*. The observer contains L Transformer (Vaswani et al., 2017) layers and E experts per layer. At every layer, the router selects or assigns probability mass to experts for each contextualized token (Fedus et al., 2022; Jiang et al., 2024). Importantly, the observer is not trained or fine-tuned for AI-generated text detection: all parameters θ remain frozen throughout our experiments.

For token x_i at layer l, let $h_i^{(l)} = H_\theta^{(l)}(x_i, x_{<i}) \in \mathbb{R}^{d_l}$ denote the contextual hidden state presented to the router. A standard router applies a learned projection

$$z_i^{(l)} = R_l\Big(h_i^{(l)}\Big) = (W_l h_i^{(l)} + b_l) \in \mathbb{R}^E, \tag{1}$$

where $z_i^{(l)}$ are the router logits. The corresponding routing probabilities are

$$p_{i,e}^{(l)} = \frac{\exp(z_{i,e}^{(l)})}{\sum_{e'=1}^{E} \exp(z_{i,e'}^{(l)})}. \tag{2}$$

Collecting router responses over all layers and token positions produces two routing traces,

$$Z_\theta(x) = [z_{i,e}^{(l)}]_{l,i,e} \in \mathbb{R}^{L\times S\times E}, \qquad P_\theta(x) = [p_{i,e}^{(l)}]_{l,i,e} \in \mathbb{R}^{L\times S\times E}. \tag{3}$$

These tensors are not additional representations trained for detection. They are native traces of how the pretrained observer allocates computation while processing the text. Our central empirical question is whether $Z_\theta(x)$ or $P_\theta(x)$ systematically differs with authorship.

Routing as an information bottleneck. The router observes a high-dimensional contextual representation $h^{(l)} \in \mathbb{R}^{d_l}$ but exposes only an E-dimensional response. Thus, even when authorship is linearly represented somewhere in the hidden state, there is no general guarantee that this direction survives the router projection.

Theorem 3.1 (Linear recoverability through the router)**.** *For a fixed layer l, let*

$$z = W_l h + b_l,$$

and consider a linear detection score in hidden space

$$s_h(h) = a^\top h + c.$$

There exists a linear score

$$s_z(z) = v^\top z + \beta$$

such that

$$s_z(W_l h + b_l) = s_h(h), \quad \textit{for every } h$$

if and only if

$$a \in \mathrm{range}(W_l^\top). \tag{4}$$

When this condition holds, any v satisfying $W_l^\top v = a$, together with $\beta = c - v^\top b_l$, exactly recovers the hidden-space score.

This states that only hidden-space directions aligned with the row space selected during language-model pretraining can be linearly preserved by router logits.

Whether authorship information happens to survive this pretrained bottleneck, and whether it survives strongly enough to support generalizable detection, is therefore an empirical question.

3.2 A Minimal Routing Readout

To study this question with as little task-specific capacity as possible, we average router responses over tokens and fit a single linear classifier . For $U_\theta(x) \in \{Z_\theta(x), P_\theta(x)\}$, we define

$$\mu_{\mathrm{mean}}(U_\theta(x)) = \mathrm{vec}\left(\frac{1}{S}\sum_{i=1}^{S} U_\theta(x)_{:,i,:}\right) \in \mathbb{R}^{LE}. \tag{5}$$

This aggregation retains layer and expert identity but does not explicitly model token order. After standardization using training-split statistics, we compute

$$\hat{y} = \sigma\big(w^\top \mathrm{Std}_{\mathrm{train}}(\mu_{\mathrm{mean}}(U_\theta(x))) + b\big), \tag{6}$$

where σ is the sigmoid function. Only (w, b) are trained with binary cross-entropy, while the observer remains frozen. For raw logits, the pre-sigmoid score is affine in the routing trace, providing a restrictive test of expert-identity-aware linear readability without nonlinear feature construction.

We distinguish four views of routing to examine how authorship information survives from router activation to expert allocation. Raw logits retain the full router response; centered logits subtract the mean over experts at each token and layer while preserving relative preferences. Routing probabilities express normalized expert preferences, whereas top-K selections retain only the identities of experts receiving computation. Together, these comparisons test whether authorship remains linearly readable in relative expert preferences and discrete expert allocation, rather than only in raw router responses.

3.3 DISSECTING THE ROUTING FOOTPRINT

After establishing whether the minimal readout works, we use additional probes to understand *why*. We study within three questions: (1) what aspects of routing carry authorship information, (2) where the information appears across model depth, and (3) whether the footprint is primarily global or depends on contextual routing dynamics.

3.3.1 WHAT ASPECTS OF ROUTING CARRY THE SIGNAL?

Global Routing Statistics. We first exam whether authorship can be explained by coarse properties of expert usage without retaining expert identities. Denote $\bar{p}^{(l)}$ the average routing probabilities over tokens, we extract statistics describing the shape of the routing distribution, including normalized entropy, normalized concentration (HHI), maximum expert probability, top-K probability mass, and the top-K boundary margin. Let $g(\cdot)$ denote this statistics map:

$$\mu_{\text{stat}}(P_\theta(x)) = \text{Concat}_{l=1}^{L}\, g\Big(\bar{p}^{(l)}\Big). \tag{7}$$

Comparing this probe against Eq. 5 distinguishes coarse routing concentration from expert-identity-specific information.

Cross-Layer Interactions. As an auxiliary structural analysis, we test whether coordinated routing changes between adjacent layers contain information beyond mean expert preferences. We first center the routing probabilities within each layer as $\tilde{p}_i^{(l)} = p_i^{(l)} - \frac{1}{S}\sum_{t=1}^{S} p_t^{(l)}$, and then compute the adjacent-layer covariance:

$$C^{(l)} = \frac{1}{S}\sum_{i=1}^{S} \tilde{p}_i^{(l)} \left(\tilde{p}_i^{(l+1)}\right)^\top, \qquad A^{(l)} = C^{(l)} \odot C^{(l)}. \tag{8}$$

We summarize the incident cross-layer covariance associated with each expert and use the resulting vector as an auxiliary probe. This analysis tests whether authorship affects not only which experts are preferred, but also how expert preferences co-vary across depth.

3.3.2 WHERE DOES THE ROUTING FOOTPRINT EMERGE?

To localize authorship information across depth, we construct both single-layer and cumulative-layer probes. For a set of layers $\mathcal{B} \subseteq \{1, \ldots, L\}$, we define

$$\mu_{\mathcal{B}}(P_\theta(x)) = \text{Concat}_{l\in\mathcal{B}} \left(\frac{1}{S}\sum_{i=1}^{S} p_i^{(l)}\right). \tag{9}$$

Setting $\mathcal{B} = \{l\}$ measures readability at an individual layer, whereas $\mathcal{B} = \{1, \ldots, l\}$ measures how detection changes as routing information accumulates through the network.

3.3.3 IS THE FOOTPRINT GLOBAL OR CONTEXT-DEPENDENT?

Our prior detector intentionally removes token order. We therefore separately investigate whether ordered routing dynamics provide information beyond the global expert-preference signal.

Context-Length Scaling. For an observation budget c, let $s_c = \min(c, S)$. We compute

$$\mu_{\text{prefix},c}(P_\theta(x)) = \text{vec}\left(\frac{1}{s_c}\sum_{i=1}^{s_c} P_\theta(x)_{:,i,:}\right). \tag{10}$$

Linear readout projector is trained for each observation length, separately. This measures both how early the routing footprint becomes detectable and how readability scales with additional context.

Contextual Position. Longer prefixes provide more router observations and include tokens processed under longer causal contexts. To disentangle these factors, we compare equal-length tokens from the prefix, random positions, and suffix of each trace. Sampling is applied only after routing traces are extracted, so every sampled token retains the contextual state produced by its original prefix.

Table 1: Main comparison on MIRAGE. Metrics: AUROC; MCC; B.Acc (balanced accuracy); TPR@5 (TPR@5%FPR). Best results are shown in **bold**. ♢ denotes a zero-shot detector, while ♠ denotes a training-based detector fitted on our same training split. Imp. reports the relative improvement of logits over the strongest baseline, computed as $(\text{new} - \text{old})/(1 - \text{old}) \times 100\%$.

	Cross-Domain				Cross-Generator				Cross-Task			
Method	AUROC	MCC	B.Acc	TPR@5	AUROC	MCC	B.Acc	TPR@5	AUROC	MCC	B.Acc	TPR@5
Statistical and zero-shot detectors												
Likelihood♢	0.5084	0.0121	0.5058	0.0787	0.5346	0.0232	0.5122	0.0890	0.5327	0.0195	0.5102	0.0878
Entropy♢	0.5646	0.1180	0.5486	0.0926	0.5969	0.1253	0.5653	0.1083	0.6138	0.1462	0.5766	0.1133
Fast-DetectGPT♢	0.6246	0.1467	0.5761	0.2240	0.6350	0.1591	0.5822	0.2292	0.6634	0.1987	0.6027	0.2717
Binoculars♢	0.7044	0.2562	0.6316	0.3286	0.7114	0.2626	0.6340	0.3286	0.7522	0.3239	0.6650	0.4108
Training-based detectors												
DetectAnyLLM♠	0.8634	0.4337	0.6453	0.4342	0.9598	0.7460	0.8892	0.8288	0.9567	0.7654	0.8955	0.7819
RepreGuard♠	0.9402	0.3610	0.6069	0.8148	0.9471	0.7385	0.8738	0.8248	0.9490	0.7566	0.8847	0.8334
SV-Detect♠	0.9251	0.6560	0.8424	0.6437	0.9668	0.7973	0.9029	0.8370	0.9795	0.8427	0.9262	0.9007
AdaDetectGPT♠	0.5700	0.0864	0.5454	0.1776	0.4298	-0.0693	0.4640	0.0697	0.6193	0.1350	0.5701	0.2423
Ours: frozen GPT-OSS-20B observer + linear routing readout												
Router mean prob.♠	0.9376	0.7000	0.8598	0.7089	0.9908	**0.9006**	**0.9588**	**0.9620**	0.9927	0.9137	0.9640	0.9689
Router mean logits♠	**0.9663**	**0.7951**	**0.9045**	**0.8442**	**0.9909**	0.8949	0.9572	0.9615	**0.9937**	**0.9185**	**0.9668**	**0.9733**
Imp.	+43.65%	+40.44%	+39.40%	+15.87%	+72.59%	+48.15%	+55.92%	+76.38%	+69.27%	+48.19%	+55.01%	+73.11%

Cross-Position Dependency. Finally, we study whether changes across token positions contain additional information. For a gap δ, we compute per-expert squared differences in raw router logits:

$$\mu_{\text{gap},\delta}(Z_\theta(x)) = \text{vec}\left(\frac{1}{S-\delta}\sum_{i=1}^{S-\delta}[Z_\theta(x)_{:,i+\delta,:} - Z_\theta(x)_{:,i,:}]^{\odot 2}\right). \tag{11}$$

We compare this representation with a synchronously shuffled-position control and with a mean-logit baseline. Any improvement that disappears after shuffling provides evidence that ordered positional relationships contribute information beyond global routing statistics.

4 Experiments

4.1 Main Results

Setup. We choose **GPT-OSS-20B** (Agarwal et al., 2025) as our primary observer MoE model, considering its model capacity, architecture novelty, and computational cost. Throughout all experiments, the observer model backbone is completely frozen. We report two minimal readouts introduced in Section 3.2: the mean routing-probability readout $\mu_{\text{mean}}(P_\theta)$ and the mean router-logit readout $\mu_{\text{mean}}(Z_\theta)$. Each readout is followed by only a single linear classifier with 769 **trainable parameters**.

Baselines. We compare against three categories of representative methods. **Statistical baselines** include *Likelihood* and *Entropy*, which directly inspect token-level predictive statistics (Gehrmann et al., 2019; Mitchell et al., 2023). **Zero-shot detectors** include *Fast-DetectGPT* (Bao et al., 2024) and *Binoculars* (Hans et al., 2024), which detect distributional irregularities through language-model outputs without task-specific training. **Training-based detectors** include *DetectAnyLLM* (Fu et al., 2025), *RepreGuard* (Chen et al., 2025b), *SV-Detect* (Vishnyakov & Gaintseva, 2026), and *AdaDetectGPT* (Zhou et al., 2026). Each baseline follows the backbone and hyper-parameter configuration of its original main experiment. AdaDetectGPT uses GPT-Neo-2.7B (Black et al., 2021) for scoring and GPT-J-6B (Wang & Komatsuzaki, 2021) as reference, following its official black-box setup. We additionally evaluate these baselines using GPT-OSS-20B as the backbone to control for backbone effects; the results are reported in Appendix. For training-based methods, we use exactly the same training split as our routing readout to ensure a fair comparison.

Dataset. Our training and evaluation data derive from MIRAGE (Fu et al., 2025), a large-scale human-AI corpus, where each AI sample is rewritten or regenerated from a corresponding human text. Such pairing reduces content and topic differences between human and AI samples, making it suitable for studying authorship signals. We construct three generalization settings by holding out

Table 2: Detection performance of different routing representations. Observer: GPT-OSS-20B. T@.1: TPR at 0.1% FPR.

		Average	
Representation	Params	AUROC	T@.1
Raw logits	769	0.9836	0.6747
Centered logits	769	0.9837	0.6681
Common logit component	25	0.9181	0.1093
Routing probabilities	769	0.9737	0.6172
Top-K selection frequencies	769	0.9673	0.4800
Sorted mean probabilities	769	0.9032	0.2431
Global routing statistics	121	0.8772	0.0736

Table 3: Predictive value of structured routing interactions. Observer: GPT-OSS-20B. T@.1: TPR at 0.1% FPR.

	Cross-layer		Temporal	
Readout	AUROC	T@.1	AUROC	T@.1
Mean	**0.9737**	**0.6172**	**0.9837**	**0.6689**
Interaction	0.9557	0.1235	0.9533	0.3276
Shuffled interaction	0.9210	0.1050	0.9585	0.3462
Mean + interaction	0.9659	0.3960	0.9758	0.5678
Mean + shuffled	0.9666	0.4704	0.9768	0.5660

different attributes between training and evaluation: ***Cross-Domain (CD)*** evaluates generalization to unseen text domains; ***Cross-Generator (CG)*** evaluates generalization to unseen AI generators; and ***Cross-Task (CT)*** evaluates generalization to unseen generation tasks. Each setting we use 500 text pairs for training, then evaluate on all remaining pairs whose held-out attribute is disjoint from the training split.

Metrics. Following DetectAnyLLM (Fu et al., 2025), the main comparison reports four metrics: AUROC, Matthews correlation coefficient (MCC), balanced accuracy (B.Acc), and true-positive rate at 5% false-positive rate (TPR@5%FPR, TPR@5). Subsequent analyses additionally report TPR@0.1%FPR (TPR@.1) to more clearly reveal performance differences under a stricter constraint.

Results. Table 1 summarizes detection performance across the three generalization settings. Mean router logits achieve AUROCs of 0.9663, 0.9909, and 0.9937, with relative improvements of 43.65%, 72.59%, and 69.27% over the strongest baselines, respectively. They outperform the strongest baseline in all 12 comparisons across metrics and settings. Mean routing probabilities also outperform all compared baselines in AUROC in the cross-generator and cross-task settings, while achieving 0.9376 in the cross-domain setting, slightly below RepreGuard (0.9402), suggesting that the signal remains accessible in normalized expert preferences.

Taken together, these results show that strong generalizable detection is attainable with a frozen observer and a 769-parameter linear readout. In the next subsection we examine which aspects of routing support this readability and how the signal varies across depth and context.

Efficiency. The routing readout combines strong detection performance with efficient inference and a small trainable parameter count. ***Inference Efficiency.*** Extracting routing traces for the entire MIRAGE corpus, containing approximately 28.0M tokens, takes only 27.7s on eight NVIDIA H200 GPUs, corresponding to approximately 0.001s per $1{,}000$ tokens. ***Parameter Efficiency.*** For GPT-OSS-20B, the complete trainable component consists of only 769 parameters in a single linear projector, while the observer itself remains completely frozen. Together, these results show that strong detection does not require either expensive detection-time statistics or a high-capacity trainable representation head once the routing footprint is directly observable.

4.2 Analysis

What carries the routing footprint? Table 2 examines which aspects of routing carry authorship information. We compare raw logits, centered logits, and the removed common component to distinguish relative expert preferences from shifts shared across experts. Routing probabilities and top-K selection frequencies test whether the signal survives normalization and discrete expert selection. Finally, sorting each layer's mean probabilities removes expert identity while preserving their values, and global routing statistics summarize distributional shape.

Centering logits leaves performance nearly unchanged, whereas the removed common component performs worse, suggesting that relative expert preferences carry authorship information. Top-K selection frequencies also carry this signal, although routing probabilities achieve higher TPR at 0.1% FPR. Expert identity also matters: sorting each layer's mean probabilities preserves their values and readout capacity but reduces average AUROC from 0.9737 to 0.9032. These results suggest that expert identities provide detection information beyond the shape of the routing distribution.

Table 4: Detection performance across nine frozen MoE observers using mean logits and mean probabilities. Metrics are averaged across three settings and seeds; T@.1 denotes TPR at 0.1% FPR.

Readout / Metric	GPT-OSS* 20B	GPT-OSS 120B	Qwen1.5 Base	Qwen1.5 Chat	Qwen3 30B-A3B	Qwen3.6 35B-A3B	DeepSeek V4-Flash	OLMoE 1B-7B	Llama-4 Scout
Head params	769	4,609	1,441	1,441	6,145	10,241	11,009	1,025	769
Logits AUROC	0.9836	0.9903	0.9552	0.9584	0.9901	**0.9945**	0.9918	0.9545	0.9626
Logits T@.1	0.6747	0.7845	0.3863	0.4246	0.8272	0.8257	**0.8655**	0.3729	0.4810
Prob. AUROC	0.9737	0.9888	0.9437	0.9450	0.9809	**0.9905**	0.9313	0.9125	0.9235
Prob. T@.1	0.6172	0.7478	0.3007	0.3240	0.5972	**0.7689**	0.0044	0.2966	0.2936

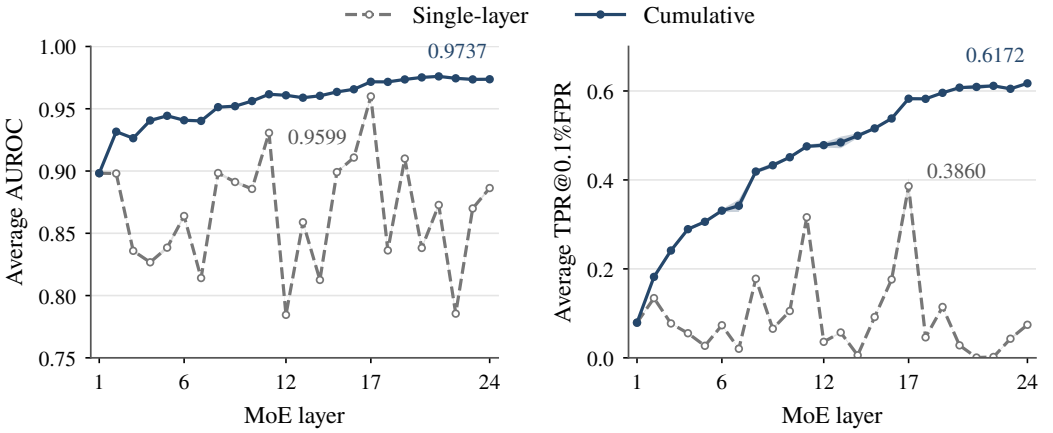


Figure 2: Detection performance across model depth. Left: AUROC; Right: TPR at 0.1% FPR. Single-layer probes use layer l; cumulative probes use layers 1 through l.

How does context shape the footprint? Figure 3 examines observation length (a) and contextual position (b) using mean routing probabilities. We first average AUROC equally across the three settings within each seed, then report the mean across seeds. On fixed cohorts of texts with at least 256 tokens, training a separate readout for each prefix length raises average AUROC from 0.7434 at 16 tokens to 0.9209 at 256 tokens. To separate token count from position, we also reuse the full-text-trained readout and average only 32 sampled token responses. Random and suffix sampling reach 0.9283 and 0.9264, respectively, compared with 0.8541 for prefix sampling. Since sampling occurs after trace extraction, each response retains its original causal context. Thus, readability varies with observation length and sampled position, suggesting that mean-pooled routing features retain contextual information.

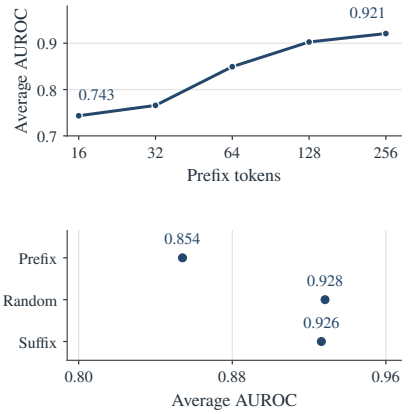


Figure 3: Context effects on AUROC.

Where does the footprint emerge across depth? Figure 2 compares single-layer readouts (33 parameters) with cumulative readouts over layers 1 through l ($32l + 1$ parameters). Single-layer AUROC varies substantially across depth, peaking at layer 17 (0.9599) and reaching 0.8863 at the final layer. Cumulative AUROC improves overall and levels off in later layers, reaching 0.9737 with all 24 layers. The full cumulative readout achieves a TPR at 0.1% FPR of 0.6172, compared with 0.3860 for the best single-layer readout. These results show that combining routing features across layers yields better detection than using any individual layer.

	Cross-layer	Temporal gap			
		16	32	64	128
CD	-1.6	-0.8	-2.2	**-2.4**	-1.9
CG	-0.4	-0.3	-0.1	-0.3	-0.4
CT	-0.3	-0.1	-0.1	-0.1	-0.7

(a) ΔAUROC (pp)

	Cross-layer	Temporal gap			
		16	32	64	128
CD	-11.9	-0.4	-15.9	-16.8	-14.2
CG	-13.1	-4.9	-4.2	-5.5	-21.1
CT	**-41.4**	-6.2	-6.6	-10.2	-15.4

(b) ΔTPR@0.1%FPR (pp)

Figure 4: Changes from adding interactions to matched mean readouts, in percentage points. Bold cells mark the largest decrease for each tested metric.

Do structured interactions add predictive value? Table 3 compares cross-layer covariance and temporal-gap features against their corresponding mean readouts: mean routing probabilities for cross-layer interactions and mean router logits for temporal interactions. We evaluate each interaction feature alone and concatenated with its mean readout, alongside shuffled controls. Shuffling breaks token alignment independently across layers for covariance features and permutes positions synchronously across layers for temporal features. Preserving cross-layer alignment raises interaction-only average AUROC from 0.9210 to 0.9557, whereas the benefit of temporal order varies across settings and gaps. Figure 4 shows three-seed mean paired changes in percentage points after adding each interaction: neither AUROC nor TPR at 0.1% FPR improves in any setting, including at each of the four temporal gaps. Temporal comparisons use a common test cohort of texts with at least 129 tokens and a mean-logit baseline matched to each gap's training population. These results favor the mean readout over the tested concatenations, which increase the parameter count from 769 to 1,537.

How much detection capability survives the routing projection? We compare hidden-state, per-layer PCA, and native router-logit readouts across 7 frozen observers, using the same setting. PCA is fitted only on training data, and matches the router's output dimension. Averaged equally across observers and settings, router, full-hidden, and PCA readouts achieve AUROCs of 0.9750, 0.9904, and 0.9873. Details can be found in Appendix. Thus, native routing compression retains strong linearly accessible detection signals, with a measurable loss.

4.3 Generality Across MoE Observers

Table 4 compares our primary observer, with eight alternative MoE models (Agarwal et al., 2025; Team, 2024; Yang et al., 2025; Qwen Team, 2026; Xu et al., 2026; Muennighoff et al., 2025; Meta, 2025). The parameter count follows the observer's native routing dimension. We report mean logits and mean probabilities, averaging AUROC and TPR at 0.1% FPR equally across the three settings and seeds. Mean-logit AUROC ranges from 0.9545 to 0.9945 across the nine observers, showing that strong authorship readability recurs across multiple MoE families.

The strength of this signal nevertheless depends on both the observer and its readout representation, particularly under strict false-positive constraints. For DeepSeek-V4-Flash, mean logits achieve average AUROC and T@.1 of 0.9918 and 0.8655, whereas mean probabilities yield 0.9313 and 0.0044, respectively. This contrast underscores the importance of readout choice across observers.

5 Conclusion

We identified an authorship routing footprint in pretrained Mixture-of-Experts models that is exposed by the native routing responses of a frozen observer and recoverable with a minimal linear readout. With GPT-OSS-20B, a 769-parameter linear classifier over token-mean router logits, trained on only 500 human-AI pairs, achieves AUROCs of 0.9663, 0.9909, and 0.9937 on MIRAGE cross-domain, cross-generator, and cross-task evaluations, respectively, outperforming all evaluated baselines. Controlled probes show that the footprint is primarily carried by relative, expert-specific preferences rather than globally shared logit shifts, while discrete expert-selection frequencies retain a strong detection signal. Combining routing information across layers improves detection, and the footprint remains readable across different observation lengths and sampled token positions and recurs across nine frozen MoE observers. The choice of routing representation nevertheless remains important, particularly under strict false-positive constraints. Overall, these results show that native MoE routing provides a structured and inexpensive observation window for AI-generated text detection. Meanwhile, such MoE router representations may emerge during its pretraining process and influence expert specialization, which finding may help a further understanding of MoE models.

Ethics Statement

This work studies AI-generated text detection using an existing benchmark and released pretrained model checkpoints; it introduces no new data collection or interaction with human participants. AI-text detectors can support provenance analysis and the investigation of misuse, but they can also cause harm when their predictions are treated as definitive evidence of authorship. False positives are especially consequential in education, employment, publishing, and content moderation, and performance may vary across domains, languages, demographic groups, generators, and editing strategies not represented by the benchmark. The proposed detector should therefore not be used as the sole basis for punitive or other high-stakes decisions. Any deployment should include population-specific evaluation, threshold calibration, privacy- and license-respecting data handling, human review, and a meaningful process for contesting decisions. Publishing the detector may also inform attempts to evade detection; we judge that transparent evaluation and reproducible scrutiny provide the more responsible basis for understanding these limitations.

Reproducibility Statement

Sections 3 and 4 specify the method, evaluation settings, baselines, and metrics, while Appendix gives the data construction, UID-based pairing, preprocessing, model configurations, optimization, random seeds, model-selection procedure, and metric aggregation details; the remaining appendix sections document the controls and full analyses. The supplementary material provides the source code, dependency specification, experiment configurations, execution and validation scripts, CPU tests, machine-readable per-run and aggregate results, evaluation-cohort UIDs and labels, provenance metadata, and integrity checksums. It also identifies the externally hosted model checkpoints and MIRAGE data needed for full reproduction and explains which large or license-governed artifacts are not redistributed.

AI Use Statement

Generative AI tools were used to assist with language editing and polishing; to search for and identify candidate references for parts of the related-work review; to assist with writing and debugging code; and to organize and prepare the submission's supplementary material. The authors reviewed all AI-assisted text, consulted the original publications and manually verified every included citation for relevance and bibliographic accuracy, and inspected and tested AI-assisted code and checked the resulting artifacts and reported results. Generative AI was not treated as an authoritative source, did not determine the scientific claims or conclusions, and was not used to fabricate or alter experimental data or results. The authors take full responsibility for the paper, citations, code, supplementary material, and all other AI-assisted content.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Elie Antoine, Frédéric Béchet, and Philippe Langlais. Part-of-speech sensitivity of routers in mixture of experts models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6467–6474, 2025.

Stefan Arnold, Marian Fietta, and Dilara Yesilbas. Routing in sparsely-gated language models responds to context. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 15–22, 2024.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, volume 2024, pp. 24814–24836, 2024.

Guangsheng Bao, Yanbin Zhao, Juncai He, and Yue Zhang. Glimpse: Enabling white-box methods to use proprietary models for zero-shot llm-generated text detection. In *International Conference on Learning Representations*, volume 2025, pp. 51844–51865, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *Zenodo*, 2021.

Jiaqi Chen, Xiaoye Zhu, Tianyang Liu, Ying Chen, Chen Xinhui, Yiwen Yuan, Chak Tou Leong, Zuchao Li, Long Tang, Lei Zhang, et al. Imitate before detect: Aligning machine stylistic preference for machine-revised text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23559–23567, 2025a.

Xin Chen, Junchao Wu, Shu Yang, Runzhe Zhan, Zeyu Wu, Ziyang Luo, Di Wang, Min Yang, Lidia S Chao, and Derek F Wong. Repreguard: Detecting llm-generated text by revealing hidden representation patterns. *Transactions of the Association for Computational Linguistics*, 13:1812–1831, 2025b.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1280–1297, 2024.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.674. URL https://aclanthology.org/2024.acl-long.674/.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Jiachen Fu, Chun-Le Guo, and Chongyi Li. Detectanyllm: towards generalizable and robust detection of machine-generated text across domains and models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11229–11238, 2025.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations*, pp. 111–116, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qrU3yNfX0d.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. MAGE: Machine-generated text detection in the wild. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36–53, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.3. URL https://aclanthology.org/2024.acl-long.3/.

Luxu Liang and Xiang Li. Steer-to-detect: Probing hidden representations for detection of llm-generated texts. *arXiv preprint arXiv:2605.12890*, 2026.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Runheng Liu, Heyan Huang, Xingchen Xiao, and Zhijing Wu. Zero-shot detection of llm-generated text via implicit reward model. *Advances in Neural Information Processing Systems*, 38:167683–167709, 2026.

Guoxin Ma, Xiaoming Liu, Hongyang Chen, Chengzhengxu Li, Zhaohan Zhang, Shengchao Liu, Yu Lan, Cong Wang, and Chao Shen. Deer: Disentangled mixture of experts with instance-adaptive routing for generalizable machine-generated text detection. *arXiv preprint arXiv:2511.01192*, 2025.

Meta. Llama 4 Model Card, April 2025. URL https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pp. 24950–24962. PMLR, 2023.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. Olmoe: Open mixture-of-experts language models. In *International Conference on Learning Representations*, volume 2025, pp. 62061–62121, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EemtbhJOXc.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, and Nicholas Andrews. Few-shot detection of machine-generated text using style representations. In *International Conference on Learning Representations*, volume 2024, pp. 29392–29412, 2024.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, 2023.

Qwen Team. Qwen1.5-moe: Matching 7b model performance with 1/3 activated parameters", February 2024. URL https://qwenlm.github.io/blog/qwen-moe/.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1702–1717, 2024.

Mikhail Vishnyakov and Tatiana Gaintseva. Sv-detect: Ai-generated text detection with steering vectors. *arXiv preprint arXiv:2606.07313*, 2026.

Ben Wang and Aran Komatsuzaki. Gpt-j-6b: A 6 billion parameter autoregressive language model, 2021.

Yan Wang, Yitao Xu, Nanhan Shen, Jinyan Su, Jimin Huang, and Zining Zhu. The illusion of specialization: Unveiling the domain-invariant" standing committee" in mixture-of-experts models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14601–14618, 2026.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltỳnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 19(1):26, 2023.

Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S Chao. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *Advances in Neural Information Processing Systems*, 37:100369–100401, 2024.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Ruochong Xiong, Qien Li, Wangwang Lian, Yulong Wan, Hanlin Xue, Zhouxing Tan, Han Yang, Fengyu Lu, and Junfei Liu. Verifiable llm-generated text detection via projected semantic-structural distributions. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14005–14042, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=1YDeZU8Lt5.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Charles Ye, Bo Yuan, and Lee Sharkey. Polysemantic experts, monosemantic paths: Routing as control in moes. In *ICLR 2026 Workshop on Representational Alignment (Re^4-Align)*, 2026. URL https://openreview.net/forum?id=IfK2ygp25S.

Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Nenghai Yu. Text fluoroscopy: Detecting llm-generated text through intrinsic features. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 15838–15846, 2024.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Hongyi Zhou, Jin Zhu, Pingfan Su, Kai Ye, Ying Yang, Shakeel Gavioli-Akilagun, and Chengchun Shi. Adadetectgpt: Adaptive detection of llm-generated text with statistical guarantees. *Advances in Neural Information Processing Systems*, 38:89077–89118, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.