VISSCORE: LLM-GENERATED TEXT DETECTION VIA WORKSPACE CONCEPT VISIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting large language model (LLM) generated text is central to trustworthy AI, yet existing detectors, whether training-based classifiers, or statistics-based detectors, return a bare probability or score: they tell the user that a text is LLM-generated, but not *which words* make it look so. We start from an observation about the scoring model itself: it is an LLM, trained on text distributions similar to those of the generators it must detect, so its internal computation should resonate with the stylistic signature of LLM-generated text. Decoding this computation into the model's own vocabulary space exposes the reader's *workspace* and confirms the *concept-visibility hypothesis*: LLM-generated text saturates a small, human-readable "flowery prose" vocabulary, while human text saturates a casual, conversational register. We build VISSCORE on this signal: decode a single golden layer of a frozen reader, discover LLM- and human-signature vocabularies from train-split logit differentials, and score each document by how strongly its workspace activates these words. Since the score aggregates named, readable words, the evidence behind every verdict can be directly inspected and audited by a human reviewer. On DetectRL-X, spanning eight languages, four generators, and six domains, VISSCORE reaches document-level AUC of 99.8 on English and 97.4 on the 8-language mean with a Qwen3-8B reader, outperforming all 11 baselines. To further explore the potential of this signal, we fit a $\sim 10^4$-parameter classifier over simple statistics of the concept-hit sequences; by recovering the structure that the scalar score discards, it lifts the mean AUC to 98.6 and, at the deployment-relevant 1% false-positive rate, the F_1 from 80.8 to 94.1. VISSCORE also holds up in the wild: it generalizes to unseen languages, generators, and domains, needs only a few labeled examples to calibrate, and stays robust to varying text lengths, LLM-assisted editing, and paraphrasing or perturbation attacks.

1 INTRODUCTION

Large language models produce fluent, persuasive text at scale, and their misuse for fake news (Pagnoni et al., 2022), academic dishonesty (Cotton et al., 2024), and disinformation (Crothers et al., 2023), together with biases (Tjuatja et al., 2024) and hallucinations (Ji et al., 2023) in LLM-generated text (LGT), makes reliable LGT detection a requirement for trustworthy deployment. Existing detectors primarily fall into two camps (Wu et al., 2025): *training-based classifiers*, which train a neural model on labeled data, and *statistics-based methods*, which score a text by feature metrics derived from a scoring model and compare to a threshold.

Training-based classifiers generally offer higher accuracy but require large amounts of labeled data, overfit to training distributions, and provide no interpretable evidence for a flagging decision, leaving them brittle under distribution shift (Wu et al., 2024). Statistics-based methods offer better interpretability and require only a few samples to set a threshold, but they suffer from two limitations: (i) they operate at the document level, providing no localization; a reviewer cannot tell which parts of the text triggered the verdict; and (ii) feature metrics such as likelihood, rank, entropy, and perplexity ratios (Solaiman et al., 2019; Su et al., 2023; Hans et al., 2024) are scalar summaries that discard the most useful clue the reader model is computing internally. A recent line of work, RepreGuard (Chen et al., 2025), partially addresses the interpretability gap by using the raw hidden representations of a surrogate model. RepreGuard extracts a discriminative direction from hidden states and classifies by projection score (RepreScore). It is effective, but it operates at the representation-statistics level: the

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.

signal is an aggregate of hidden-state covariances or principal components over the whole sequence, not expressed in any human-readable form. It cannot be traced to the words that triggered the verdict, so the interpretability gain over a black-box classifier is limited.

Our starting point is an observation about the scoring model itself: it is an LLM, trained on text distributions similar to those of the generators it must detect, so its internal computation should resonate with the stylistic signature of LGT. We hypothesize that, once decoded into the model's own vocabulary space, this resonance surfaces as readable, word-level signatures rather than an opaque scalar statistic: a set of words the reader activates systematically differently on LGT than on HWT. We call this the *concept-visibility hypothesis*. Decoding a frozen reader on $9{,}600$ training pairs confirms it (Figure 1): LGT saturates a small "flowery prose" vocabulary, whereas HWT saturates a casual, conversational register that LGT enters markedly less. The decoded concepts are thus simultaneously the detection feature and the evidence shown to a reviewer; the explanation is not post-hoc. The readout that makes this observation possible is vocabulary-space decoding, which originates with the logit lens (nostalgebraist, 2020): read out the distribution the model would emit as if layer l were the final layer, rendering content the model computes internally but never emits as output readable as words. We refer to the intermediate content exposed this way as the reader's *workspace*, and to the decoded per-layer distributions as *workspace logits*. We adopt the plain logit lens throughout this paper and compare the effect of different decoders in §5.3.

Building on this observation, we propose VISSCORE, an LGT detector that scores a document by how strongly a frozen reader's decoded workspace activates LLM- versus human-signature words, so that the words driving each verdict remain directly readable. Concretely, we decode the residual stream of the reader at a single *golden layer*, chosen once by a systematic layer scan, detailed in §5.4, and shared across all languages, to obtain per-position workspace logits. From these we discover LLM- and human-signature concept vocabularies by a per-token mean positive logit differential under false-discovery-rate (FDR) control on the train split, and compute VISSCORE, the difference between LLM-concept and human-concept hit rates at the golden layer, compared against a threshold. VISSCORE is already a strong statistics-based detector on its own, exceeding per-language fine-tuned XLM-RoBERTa-base without any classifier training, and a $\sim 10^4$-parameter classifier over simple statistics of the concept-hit sequences boosts it further.

Our contributions are as follows: (1) We propose and validate the concept-visibility hypothesis: a frozen reader's decoded workspace separates LGT from HWT through small, human-readable signature vocabularies (§3.1). The signal resides in intermediate layers rather than the final layer (§5.4) and is not tied to a specific decoder (§5.3); (2) We build VISSCORE, a statistics-based detector that fits nothing beyond a threshold, on this signal (§3). Because the signal is a set of named words, every verdict carries readable evidence a reviewer can directly inspect, with the same activations localizing to spans when finer granularity is needed (Appendix A); (3) We evaluate VISSCORE on DetectRL-X across eight languages, four generators, and six domains, reaching document-level AUC 99.8 on English and 97.4 on the 8-language mean and outperforming all 11 baselines; a $\sim 10^4$-parameter classifier over the concept-hit statistics lifts the mean to 98.6 (§5.1). VISSCORE further shows the best out-of-distribution (OOD) generalization across languages, generators, and domains (§5.2), together with strong sample efficiency (§5.5) and reliability in the wild (§5.6).

2 RELATED WORK

Statistics-based detection. Statistics-based detectors score each document with a fixed statistic or projection and require no classifier training. One family derives the score from output logits: **Entropy**, **Log-Likelihood**, and **Log-Rank** (Solaiman et al., 2019), combined into the **LRR** by Su et al. (2023); **DetectGPT** (Mitchell et al., 2023), its efficient variant **Fast-DetectGPT** (Bao et al., 2024), and **Binoculars** (Hans et al., 2024) exploit the curvature or ratios of log-probabilities. A second family derives the statistic from internal representations: **RepreGuard** (Chen et al., 2025) extracts a discriminative PCA direction from surrogate-model representations and scores by projection (RepreScore). Both families return scalar statistics: these feature scores are opaque and carry no readable evidence that a reviewer can directly inspect.

Training-based detection. Training-based detectors fit a classifier to labeled LGT/HWT data: fine-tuning a multilingual encoder such as **XLM-RoBERTa** (Conneau et al., 2020) per language

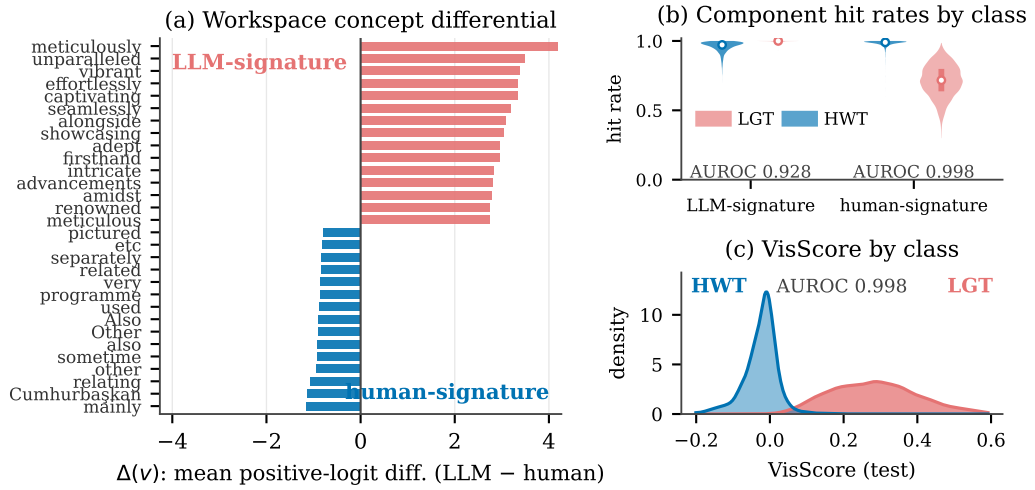


Figure 1: Hypothesis validation (Qwen3-8B reader, DRX English). (a) Per-token concept differential $\Delta(v)$ of Eq. 3 at the golden layer $l^\star$=22 from 9,600 train pairs: LLM-signature concepts (light red) are "flowery prose" tokens. Human-signature concepts (blue) are casual, conversational vocabulary. (b) Document-level hit rates per concept set in LGT vs. HWT (violin: distribution; white dot: median). (c) Test-set VISSCORE distributions of LGT and HWT (AUC 99.8).

remains the standard reference point. **SeqXGPT** (Wang et al., 2023) formulates sentence-level detection as sequence labeling over token-level log-probability sequences; **BiScope** (Guo et al., 2024) adds a backward cross-entropy from the output logits to the immediately preceding input token, exploiting LLMs' memorization of preceding tokens; **DivEye** (Basani & Chen, 2026) trains on surprisal-diversity features, combining distributional moments and first/second-order temporal differences of the token-surprisal sequence. These detectors tend to overfit to their training distribution and provide no interpretable evidence for individual decisions.

Vocabulary-space decoding. A line of interpretability work decodes a frozen model's intermediate hidden states into vocabulary space (nostalgebraist, 2020; Belrose et al., 2023; Gurnee et al., 2026); §3.1 reviews these decoders. They were developed to interpret a model's own computation; we instead repurpose vocabulary-space decoding as a *signal-extraction* mechanism for LGT detection, where the decoded concepts serve simultaneously as the discriminative feature and as human-readable evidence, so the score and its evidence are the same object.

3 VISSCORE FOR LLM-GENERATED TEXT DETECTION

3.1 WORKSPACE DECODING AND THE CONCEPT-VISIBILITY HYPOTHESIS

We define the reader's *workspace* as its intermediate computation (the residual-stream state $\boldsymbol{h}_l(x_i)$ at every position i and layer l) and *workspace decoding* as the operation that maps these states back into vocabulary space so that their content becomes inspectable as words. Formally, given a transformer reader model $\mathcal{M}$ with L layers and hidden dimension d, a *vocabulary-space decoder* $\mathcal{D}_l$ maps the layer-l residual stream $\boldsymbol{h}_l(x)$ toward the final layer's coordinates. Projecting through the unembedding $\boldsymbol{W}_U$ yields *workspace logits*:

$$\hat{\boldsymbol{z}}^{(l)}(x) = \boldsymbol{W}_U\,\mathcal{D}_l\big(\boldsymbol{h}_l(x)\big) + \boldsymbol{b}_U, \qquad \hat{\boldsymbol{y}}^{(l)}(x) = \mathrm{softmax}\big(\hat{\boldsymbol{z}}^{(l)}(x)\big), \tag{1}$$

i.e., $\hat{\boldsymbol{y}}^{(l)}(x)$ is the output distribution as if the residual stream at layer l were the final layer, and $\hat{\boldsymbol{z}}^{(l)}(x)$ its pre-softmax counterpart. At each position i and layer l, the workspace logits $\hat{\boldsymbol{z}}_i^{(l)} \in \mathbb{R}^{|\mathcal{V}|}$ reveal what vocabulary the reader is "thinking about" at that position and layer. Workspace logits are therefore the reader's predictions, not the text's surface content: a concept can saturate the workspace without ever occurring as a token, and a human text quoting flowery vocabulary still

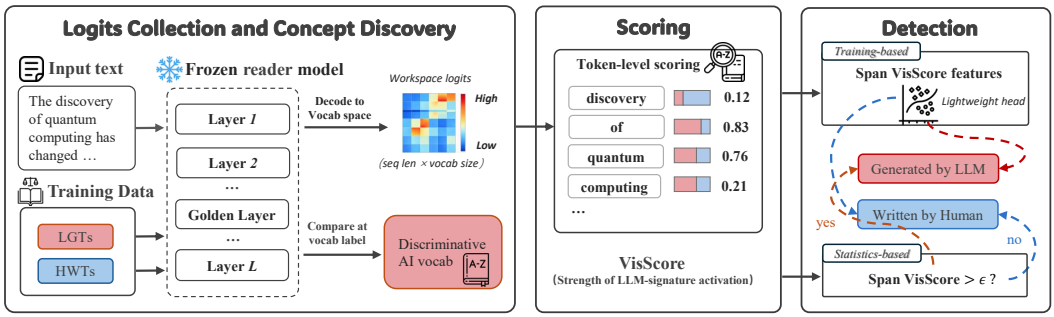


Figure 2: Overview of VISSCORE. *Left*: a frozen reader is decoded at the golden layer into per-position workspace logits, and comparing LGT and HWT training pairs at the vocabulary level yields the LLM- and human-signature concept vocabularies. *Middle*: VISSCORE scores a text by the difference of the two concept hit rates. *Right*: detection flags the text by a train-fitted threshold, or routes visibility-facet features through a lightweight head. Words and values are schematic.

yields casual-register predictions elsewhere. The signal measures how the reader expects the text to continue, so it is not reducible to surface counts of "LLM-register" words (Kobak et al., 2025).

Several decoders can instantiate $\mathcal{D}_l$: the identity map (logit lens; nostalgebraist, 2020), learned per-layer affine probes (tuned lens; Belrose et al., 2023), or the expected input-independent Jacobian $\boldsymbol{J}_l \approx \mathbb{E}[\partial \boldsymbol{h}_{\text{final}} / \partial \boldsymbol{h}_l]$ (Jacobian lens; Gurnee et al., 2026). We adopt the plain logit lens, the identity transport $\mathcal{D}_l(\boldsymbol{h}_l) = \boldsymbol{h}_l$, as the default decoder: it is the simplest possible instantiation and requires no fitting of any kind, so the entire pipeline trains nothing beyond the concept vocabulary and a threshold. The signal is not tied to this choice: the Jacobian lens performs on par, as §5.3 shows.

The premise of VISSCORE is the concept-visibility hypothesis: a frozen LLM, when reading a text, activates a systematically different set of concepts on LGT than on HWT; decoded into vocabulary space, this difference surfaces as readable, word-level signatures rather than an opaque scalar statistic.

Figure 1 validates the hypothesis empirically. When we process $9{,}600$ LGT/HWT pairs through the Qwen3-8B reader, the per-token workspace differential $\Delta(v)$ of Eq. 3 at layer 22 surfaces two highly interpretable vocabularies in Figure 1a; Table 4, Appendix A, lists the top-100 tokens of each side with their differentials. The LLM-signature set consists of canonical "flowery prose": *impactful, effortlessly, meticulously, seamlessly, captivating, burgeoning, insightful, transformative*. The human-signature set consists of casual, functional and narrative vocabulary: *mainly, also, used, programme, totally, actually, probably, organisation*. Their test-set statistics in Figure 1b reveal an asymmetric structure. The LLM-signature register is shared rather than exclusive: HWT activates it almost as heavily, at a median hit rate of 97%, but human usage varies widely across domains, peaking in formal writing such as academic prose, whereas LGT saturates the register in every document, at a median of 100%. LLM-generated writing converges on a single elevated register, and human writing samples many. The human-signature register is the mirror image and closer to genuine exclusion: HWT saturates it at a median of 99%, while LGT enters it markedly less, at a median of 72%, and this channel alone is the stronger detector, with AUC 99.8 versus 92.8. The absence of casual vocabulary is more diagnostic than the presence of flowery vocabulary. VISSCORE's difference exploits both directions at once. The separation is broad rather than point-localized, as §5.4 and Figure 3 show, so VISSCORE decodes only the golden layer. The resulting score separates the two classes on held-out data: test-set VISSCORE distributions of LGT and HWT form two well-separated modes with AUC 99.8 in Figure 1c.

3.2 THE VISSCORE PIPELINE

Based on this hypothesis, we propose VISSCORE for detecting LGT from the concept activations in the reader's workspace. Figure 2 gives an overview of the pipeline, which separates a one-time offline stage on the train split from detection time: workspace-logit collection and concept discovery run once on paired training texts (left); a new text is then decoded at the same golden layer, scored (middle), and flagged (right), with no further fitting.

Workspace Logits Collection and Concept Discovery. We introduce a small training set $\{(\mathcal{T}^i_{\text{LGT}}, \mathcal{T}^i_{\text{HWT}})\}_{i=1}^{N}$ of LGT/HWT pairs and use the reader $\mathcal{M}$ as an "observer". For each text $\mathcal{T} = \{t_1, \ldots, t_n\}$, we process it through $\mathcal{M}$, decoding the layer set $\mathcal{L}$: for each token t_j and layer $l \in \mathcal{L}$, workspace logits $\hat{\boldsymbol{z}}_j^{(l)} \in \mathbb{R}^{|\mathcal{V}|}$ follow from Eq. 1. The complete workspace activation of $\mathcal{T}$ is:

$$\mathcal{W}(\mathcal{T}) = \{\hat{\boldsymbol{z}}_j^{(l)} \mid j \in [1, n],\ l \in \mathcal{L}\}. \tag{2}$$

We decode a set of layers $\mathcal{L}$ and use workspace size K=500; the default throughout this paper is the single golden layer $\mathcal{L} = \{l^\star\}$, $l^\star$=22. Any layer set is admissible: the full 36-layer scan of Figure 3 shows that one well-chosen layer nearly saturates the signal and already exceeds the all-layer average, so no other layer is decoded; K and $|\mathcal{C}|$ are ablated in §5.4.

Rather than hand-designing the LLM-signature vocabulary, we let the data reveal it. For each vocab token $v \in \mathcal{V}$, we compute the per-token mean positive logit differential at the golden layer:

$$\Delta(v) = \frac{1}{|\mathcal{D}_{\text{ai}}|} \sum_{x \in \mathcal{D}_{\text{ai}}} \sum_i \max\big(\hat{z}_{i,v}^{(l^\star)},\, 0\big) \;-\; \frac{1}{|\mathcal{D}_{\text{hu}}|} \sum_{x \in \mathcal{D}_{\text{hu}}} \sum_i \max\big(\hat{z}_{i,v}^{(l^\star)},\, 0\big), \tag{3}$$

where $\mathcal{D}_{\text{ai}}$ and $\mathcal{D}_{\text{hu}}$ are the LLM-generated and human text sets from the training split. After filtering unreadable tokens (pure punctuation and tokens with no alphabetic characters; readable BPE fragments are kept), we test each candidate token for a stable differential via a pair-level bootstrap (resampling LGT/HWT pairs) with Benjamini–Hochberg FDR control at 5%. The top-500 surviving tokens by Δ form the LLM-signature concept set $\mathcal{C}_{\text{ai}}$, and the top-500 by $-\Delta$ the human-signature set $\mathcal{C}_{\text{hu}}$; FDR control can leave fewer than 500 survivors on a side, so $|\mathcal{C}|$ is a cap rather than a fixed size (English: 500 LLM vs. 325 human tokens; Portuguese: 500 vs. 46). The discovered vocabularies are interpretable and analyzed with Figure 1a in §3.1.

Workspace Concept Scoring. At detection time, a new text $\mathcal{T}$ passes through the same frozen reader and is decoded at the golden layer $l^\star$. We compute the concept hit rate, the fraction of positions whose workspace top-K contains at least one token from $\mathcal{C}_{\text{ai}}$ (resp. $\mathcal{C}_{\text{hu}}$). VISSCORE is their difference:

$$\text{VISSCORE}(\mathcal{T}) = \text{hit-rate}^{(l^\star)}_{\mathcal{C}_{\text{ai}}}(\mathcal{T}) - \text{hit-rate}^{(l^\star)}_{\mathcal{C}_{\text{hu}}}(\mathcal{T}). \tag{4}$$

Detection. The score becomes a verdict in one of two ways: **Statistics-based.** A single operating threshold τ is fitted on the train split, by either F_1 maximization (τ_{F_1}) or the FPR$\leq 1\%$ constraint ($\tau_{1\%}$), and a text is flagged as LGT if $\text{VISSCORE}(\mathcal{T}) \geq \tau$. **Training-based.** This route restores the structure the scalar score discards: rather than compressing the per-position hit sequences into one number, it describes them with a compact set of *visibility-facet* features, simple statistics over the two concept-hit sequences. Six per-side statistics answer three questions: the hit rate measures how often the concept set is hit; the lag-1 autocorrelation, maximal run length, and burst fraction measure whether the hits cluster; and the center of mass and the first-/second-half trend measure where the hits fall. Three further rates compare the two sides position by position: LLM-only, human-only, and joint hits. The resulting $6 \times 2 + 3 = 15$ features form the vector $\phi(\mathcal{T}) \in \mathbb{R}^{15}$, on which a LightGBM classifier of $\sim 10^4$ parameters, trained on the same train split, predicts $p(\text{LGT} \mid \mathcal{T})$. Appendix B gives the exact definitions. Sequence length is deliberately excluded to avoid length leakage.

Readable Concept Evidence. VISSCORE produces word-readable evidence: for any span, we decode its workspace top-K tokens and read off exactly which concepts drove the score. Figure 7 in Appendix A illustrates this with a matched pair of abstracts on the same topic. In the LLM-generated text, every span saturates the LLM-signature register with concepts such as *significantly*, *dynamic*, and *insights*, while the human-signature register lags behind. In the human-written abstract, the workspace visits the same LLM-signature concepts just as densely, but human-signature words such as *associated*, *corresponding*, and *strictly* keep pace and cancel the gap. A reviewer can thus audit each verdict span by span, down to concept counts.

4 EXPERIMENTAL SETUP

Datasets. Our primary benchmark is **DetectRL-X (DRX)** (Wu et al., 2026): 72k records covering eight languages (English, Portuguese, French, Spanish, Chinese, German, Arabic, and Russian), four

Table 1: Main results on DetectRL-X (%): per-language AUC (left) and 8-language mean AUC/F_1/F_1@1% (right), all on the official test splits. Bold (underlined): best (second-best) per column within each block; ties are bolded.

	Per-language AUC								8-language Mean		
Method	English	Portuguese	French	Spanish	Chinese	German	Arabic	Russian	AUC	F_1	F_1@1%
Statistics-based											
Fast-DetectGPT (Bao et al., 2024)	56.3	63.6	65.5	63.7	52.2	67.5	63.8	74.0	63.3	66.9	11.2
Binoculars (Hans et al., 2024)	56.3	64.7	67.4	65.6	52.3	69.6	66.5	77.5	65.0	67.4	13.3
LogRank	71.3	72.2	72.1	70.6	64.0	70.3	68.2	57.5	68.3	65.8	20.3
LRR (Su et al., 2023)	78.2	86.4	85.5	88.6	51.3	85.4	75.3	74.8	78.2	74.1	19.5
Entropy	78.7	87.1	84.7	87.7	61.0	84.8	75.0	67.6	78.3	74.4	21.9
Log-Likelihood	83.4	90.5	90.9	91.1	65.3	91.0	84.9	80.2	84.7	77.6	40.0
RepreGuard (Chen et al., 2025)	97.8	**97.6**	96.3	95.2	95.7	83.6	88.3	84.2	92.3	85.3	70.2
VISSCORE (ours)	**99.8**	97.5	**98.0**	**97.5**	**98.2**	**96.3**	**98.4**	**93.4**	**97.4**	**91.9**	**80.8**
Training-based											
XLM-RoBERTa-base	**99.8**	96.6	98.9	**98.6**	**99.6**	97.2	96.2	88.9	97.0	**95.0**	59.6
BiScope (Guo et al., 2024)	99.4	97.2	97.4	97.8	98.8	96.6	94.4	91.8	96.7	90.4	91.0
SeqXGPT (Wang et al., 2023)	93.1	96.3	95.7	97.5	95.4	96.3	93.9	92.4	95.1	86.5	70.3
DivEye (Basani & Chen, 2026)	90.9	92.6	92.8	93.5	90.3	91.5	90.8	86.6	91.1	83.3	82.9
VISSCORE + LGBM (ours)	**99.8**	**98.9**	**99.0**	98.4	98.5	**98.3**	**98.9**	**96.7**	**98.6**	94.4	**94.1**

generators (deepseek-v3, gpt-4o, gemini-2.5-flash, and qwen-max), and six writing domains (News, Wiki, Webtext, SEO, Novel, and Academic), with an official train/test split per language.

Reader (surrogate) model. Following DetectRL-X (Wu et al., 2026), we instantiate the surrogate reader as **Qwen3-8B** (36 layers, d=4096), the primary reader for all experiments. The remaining configuration ($l^\star$=22, $|\mathcal{C}|\leq$500 per side per language, K=500) is as defined in §3.2; span size S=32 is used only for the appendix evidence visualization in Figure 7.

Baselines. We evaluate seven statistics-based detectors with the same Qwen3-8B reader on identical DRX splits: Log-Likelihood, Entropy, LogRank, LRR (Su et al., 2023), Binoculars (Hans et al., 2024), Fast-DetectGPT (Bao et al., 2024), and RepreGuard (Chen et al., 2025). For training-based baselines, we evaluate fine-tuned XLM-RoBERTa-base, BiScope (Guo et al., 2024), SeqXGPT (Wang et al., 2023), and DivEye (Basani & Chen, 2026) from their official implementations and classifier configurations on the identical splits. Appendix C details the configuration of every detector.

Metrics. We report AUC as the threshold-free primary criterion, F_1 at the F_1-maximal threshold, and F_1 at a false-positive rate capped at 1% (F_1@1%), which matters when falsely flagging human writing is costly. Both thresholds are fitted on the train split and applied unchanged to the test split, so all reported numbers are test-split values.

5 RESULTS

5.1 MAIN RESULTS ON DETECTRL-X

Among the statistics-based detectors in Table 1, VISSCORE is the strongest: mean AUC 97.4, mean F_1 91.9, and F_1@1% 80.8 are all the best in the group, leading the strongest baseline, RepreGuard, by 5.1 AUC points; it ranks first on seven of the eight per-language AUC columns, behind RepreGuard only on Portuguese by 0.1, and its largest margin, 12.7 points on German, falls where hidden-state projection degrades most. VISSCORE even leads the strongest training-based entry, per-language fine-tuned XLM-RoBERTa-base, on AUC, 97.4 to 97.0, and decisively on F_1@1%, 80.8 to 59.6, conceding only 2.1 points of max-F_1. VISSCORE+LGBM in the bottom block removes this last gap, attaining the best training-based results at mean AUC 98.6, mean F_1 96.1, and F_1@1% 94.1 with only $\sim 10^4$ parameters, four orders of magnitude fewer than XLM-RoBERTa-base, while inheriting VISSCORE's concept-level interpretability, which a fine-tuned encoder cannot provide.

Table 2: OOD generalization (%): AUC / F_1 / F_1@1%; cross-generator and cross-domain rows are English folds. Blue (red) subscripts mark gains (losses) vs. RepreGuard / XLM-R, computed from unrounded values, so they can differ by 0.1 from the displayed, rounded entries.

	Statistics-based						Training-based					
	RepreGuard			VisScore			XLM-R			VisScore+LGBM		
Setting	AUC	F_1	F_1@1%	AUC	F_1	F_1@1%	AUC	F_1	F_1@1%	AUC	F_1	F_1@1%
In-distribution (En.)	97.8	91.8	81.7	99.8$_{+2.0}$	98.3$_{+6.5}$	97.5$_{+15.8}$	99.8	98.6	93.8	99.8$_{+0.0}$	98.4$_{-0.2}$	98.4$_{+4.6}$
Cross-language												
Arabic	91.2	67.6	74.3	96.7$_{+5.5}$	85.7$_{+18.1}$	86.5$_{+12.2}$	64.8	18.7	0.0	97.0$_{+32.2}$	86.4$_{+67.7}$	87.1$_{+87.1}$
Chinese	82.8	7.8	2.3	96.9$_{+14.2}$	90.5$_{+82.8}$	80.5$_{+78.2}$	63.7	15.5	6.0	96.7$_{+32.9}$	89.7$_{+74.3}$	90.2$_{+84.2}$
English	96.2	85.4	59.9	98.5$_{+2.2}$	83.3$_{-2.2}$	87.0$_{+27.1}$	96.6	71.1	59.4	97.2$_{+0.5}$	84.4$_{+13.3}$	78.6$_{+19.1}$
French	96.2	80.1	86.1	96.2$_{-0.1}$	88.5$_{+8.4}$	82.8$_{-3.3}$	92.1	35.8	3.7	96.9$_{+4.9}$	91.0$_{+55.2}$	90.8$_{+87.1}$
German	87.0	72.6	59.4	95.8$_{+8.8}$	87.2$_{+14.6}$	82.9$_{+23.6}$	88.4	37.6	0.5	96.7$_{+8.3}$	89.7$_{+52.2}$	90.2$_{+89.7}$
Portuguese	97.1	87.0	81.3	95.8$_{-1.3}$	87.2$_{+0.2}$	85.8$_{+4.4}$	96.0	52.8	8.5	97.4$_{+1.5}$	91.7$_{+38.9}$	91.9$_{+83.3}$
Russian	90.4	71.2	72.8	95.0$_{+4.6}$	84.6$_{+13.4}$	69.5$_{-3.3}$	87.8	3.2	0.0	96.4$_{+8.6}$	86.6$_{+83.4}$	84.7$_{+84.7}$
Spanish	96.1	69.4	78.2	96.7$_{+0.6}$	85.5$_{+16.1}$	85.1$_{+7.0}$	95.0	55.7	10.6	96.4$_{+1.4}$	89.2$_{+33.4}$	89.4$_{+78.8}$
Mean	92.1	67.6	64.3	96.5$_{+4.3}$	86.6$_{+18.9}$	82.5$_{+18.2}$	85.5	36.3	11.1	96.8$_{+11.3}$	88.6$_{+52.3}$	87.8$_{+76.7}$
Cross-generator												
DeepSeek-V3	97.8	91.5	82.5	99.7$_{+2.0}$	98.0$_{+6.6}$	97.1$_{+14.6}$	99.0	95.5	82.6	99.8$_{+0.8}$	97.2$_{+1.8}$	98.4$_{+15.8}$
Gemini-2.5-Flash	95.9	88.1	70.6	99.9$_{+4.0}$	98.5$_{+10.3}$	98.5$_{+27.9}$	98.3	97.8	43.6	99.9$_{+1.6}$	98.5$_{+0.8}$	98.3$_{+54.7}$
GPT-4o	98.3	92.9	80.5	99.8$_{+1.5}$	99.3$_{+6.3}$	98.8$_{+18.3}$	99.5	98.7	85.2	99.9$_{+0.4}$	99.0$_{+0.4}$	98.9$_{+13.7}$
Qwen-Max	96.6	90.0	77.5	99.4$_{+2.7}$	96.7$_{+6.7}$	95.1$_{+17.6}$	99.5	97.3	94.1	99.5$_{+0.0}$	96.1$_{-1.2}$	97.5$_{+3.4}$
Mean	97.2	90.6	77.8	99.7$_{+2.6}$	98.1$_{+7.5}$	97.4$_{+19.6}$	99.1	97.3	76.4	99.8$_{+0.7}$	97.7$_{+0.4}$	98.3$_{+21.9}$
Cross-domain												
Academic	99.7	97.8	88.1	99.9$_{+0.1}$	98.2$_{+0.4}$	98.8$_{+10.7}$	98.8	95.7	87.3	99.7$_{+1.0}$	98.5$_{+2.8}$	97.5$_{+10.2}$
News	99.3	94.4	72.8	100.0$_{+0.7}$	99.3$_{+4.9}$	97.6$_{+24.8}$	100.0	99.5	93.0	100.0$_{+0.0}$	98.6$_{-0.9}$	99.1$_{+6.0}$
Novel	96.4	91.7	54.3	99.6$_{+3.2}$	96.4$_{+4.7}$	95.7$_{+41.4}$	96.3	93.0	84.9	99.5$_{+3.2}$	96.6$_{+3.7}$	95.2$_{+10.3}$
SEO	99.7	69.9	94.6	99.6$_{-0.1}$	97.2$_{+27.3}$	97.2$_{+2.7}$	98.4	83.3	12.2	99.0$_{+0.6}$	97.1$_{+13.8}$	95.5$_{+83.3}$
Webtext	97.2	71.5	81.0	99.9$_{+2.7}$	97.7$_{+26.2}$	98.8$_{+17.8}$	93.2	90.2	20.7	99.9$_{+6.7}$	98.2$_{+8.1}$	95.9$_{+75.2}$
Wiki	96.7	89.2	60.8	99.9$_{+3.2}$	95.6$_{+6.3}$	94.0$_{+33.2}$	100.0	98.1	75.5	100.0$_{\pm 0.0}$	94.0$_{-4.1}$	98.2$_{+22.7}$
Mean	98.2	85.7	75.3	99.8$_{+1.6}$	97.4$_{+11.6}$	97.0$_{+21.7}$	97.8	93.3	62.3	99.7$_{+1.9}$	97.2$_{+3.9}$	96.9$_{+34.6}$

5.2 OOD Generalization

Table 2 probes OOD generalization along three axes, with full matrices in Appendix D: cross-language, training on one language's clean split and evaluating on the other seven; cross-generator and cross-domain, holding out one generator or writing domain and re-discovering the concepts on the remaining data. All trained parameters and thresholds are fit on the training split and never calibrated on the OOD test split. VisScore transfers almost intact: the cross-language mean AUC drops by only 0.9 points (97.4 $\rightarrow$ 96.5), and unseen generators and domains cost at most 0.4 points; VisScore+LGBM retains most of its advantage in all three settings. What transfers is the discovered concept set itself, not any trained component: the workspace appears to encode the generators' stylistic signatures rather than language- or domain-specific surface cues. The baselines, by contrast, degrade through the parts they train: XLM-RoBERTa's fine-tuned weights transfer only partially, its F_1@1% collapsing to 11.1 cross-language and 62.3 cross-domain, and RepreGuard's fitted projection direction preserves the ranking but drifts in scale, so thresholds fit on the training split no longer apply and its F_1@1% falls to 23.1.

5.3 Is the Signal Tied to the Decoder?

VisScore treats vocabulary-space decoding as a generic signal-extraction step, so we test whether the signal depends on the decoder $\mathcal{D}_l$: Table 3 re-runs the entire pipeline with the logit lens replaced by two fitted alternatives, a Jacobian lens (Gurnee et al., 2026) and a tuned lens (Belrose et al., 2023). The Jacobian lens stays close, mean AUC 96.2 versus 97.4, and the two transports discover overlapping vocabularies, sharing 24 of the top-100 concepts on average: the discriminative signal is a property

Table 3: Decoder ablation on DRX at the golden layer (Qwen3-8B, 8-language mean).

Decoder $\mathcal{D}_l$	AUC	F_1	F_1@1%
Logit lens	**97.4**	**91.9**	**80.8**
Jacobian lens	96.2	90.1	67.8
Tuned lens	89.2	81.6	42.8

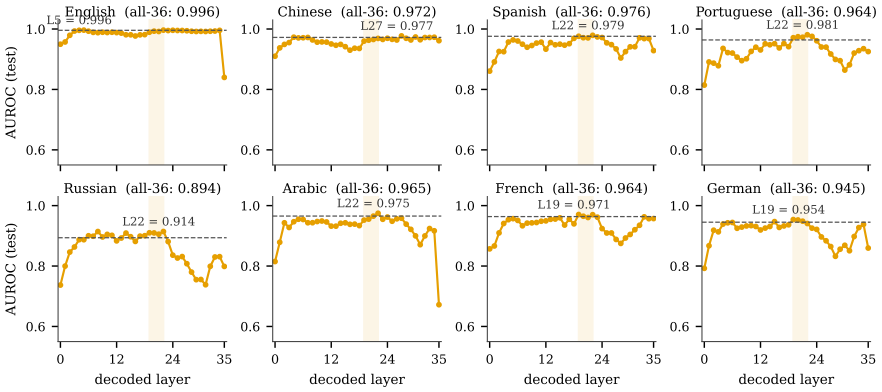


Figure 3: 36-layer scan on all 8 DRX languages (logit lens, K=500, $|\mathcal{C}|$=100; concepts discovered on the train split, AUC reported on the test split). Shaded band: layers 19–22; dashed: all-layer average.

of vocabulary-space decoding itself, not of a particular transport. The tuned lens instead collapses to 89.2 mean AUC, its vocabulary nearly disjoint from the logit lens's with only 5 of the top-100 shared, indicating that the signal resides in the raw geometry of the unembedding, not in a re-projection fitted on generic text to match final-layer distributions. We retain the logit lens: it requires no fitting and is empirically stronger at low false-positive rates, 80.8 versus 67.8 F_1@1%.

5.4 Ablations: Layers, K, and $|\mathcal{C}|$

Figure 3 ablates the decoded layer and Figure 4 the workspace size K and vocabulary size $|\mathcal{C}|$, the two remaining design choices. (i) *Layer choice is robust, and the signal is not in the final layer.* Layers 19–22 lie within 3 points of each language's optimum, and the fixed $l^\star$=22 attains the best 8-language mean AUC, 96.6, within 0.3 points of per-language oracle tuning; the final layer instead collapses everywhere. The signal resides in intermediate computation, before the reader specializes for next-token prediction, so we decode only the language-agnostic $l^\star$=22. (ii) *The vocabulary must be selective, and K and $|\mathcal{C}|$ must scale together.* Unfiltered, both sides share one vocabulary and every language falls to chance (50); at $|\mathcal{C}|$=100, no K recovers the peak (best 97.0). Only joint scaling improves, peaking at the default K=500, $|\mathcal{C}|$=500, mean AUC 98.1 on symmetric top-$|\mathcal{C}|$ vocabularies (the FDR-selected default of Table 1 attains 97.4): a wide window over a small vocabulary collects spurious hits, a narrow window starves a large one, and per-language tuning adds only 0.3 points.

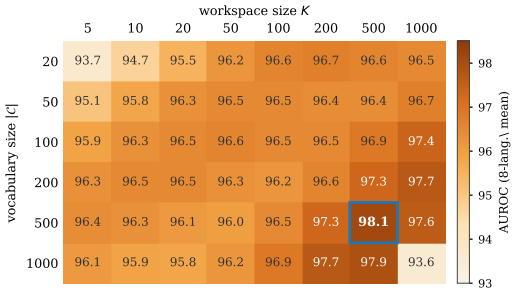


Figure 4: $K \times |\mathcal{C}|$ grid at the golden layer (AUC, 8-language mean, test split, symmetric top-$|\mathcal{C}|$ vocabularies). Boxed cell: default K=500, $|\mathcal{C}|$=500.

5.5 Sample Efficiency

VisScore exhibits strong few-shot behavior. Figure 6 varies the number of labeled pairs, sampling one pair per generator–domain cell with the concept vocabulary fixed. VisScore is essentially indifferent: AUC stays at 99.8 down to 24 pairs, and even its deployment thresholds hold, F_1@1% 96.4. VisScore+LGBM needs slightly more: from 24 pairs it reaches AUC 97.2 but its thresholds wobble, and at 1,200 pairs and beyond it matches the full-data 99.8. The baselines bracket the spectrum: RepreGuard shares the near-zero-shot behavior, 95.7 from 24 pairs, but plateaus at its full-data 97.8; XLM-RoBERTa is data-hungry, 64.0 from 24 pairs and above 99 only from 2,400 pairs. Supervision in VisScore is thus spent calibrating thresholds and a 10^4-parameter head, not building the score.

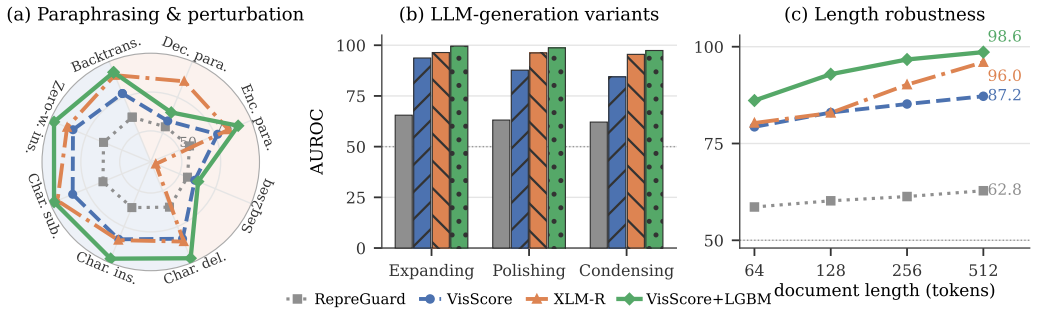


Figure 5: Robustness (AUC) on the full 8-language DRX, methods deployed as trained on clean text. (a) Paraphrasing attacks (orange half) and perturbation attacks (blue half), §5.6; (b) LLM-generation variants, Appendix E; (c) length truncation, §5.6. Dotted lines in (b, c): chance.

5.6 Reliability in the Wild

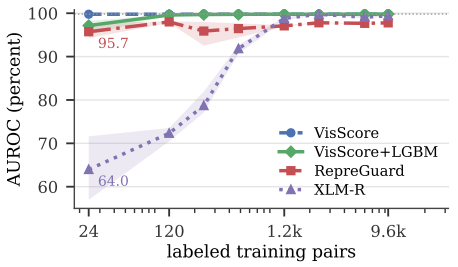


Figure 6: Sample efficiency (AUC, DRX English): test AUC vs. labeled training pairs, with VisScore+LGBM, RepreGuard, and XLM-RoBERTa refit on the identical subsets. Bands: 3-seed range.

Paraphrasing and perturbation attacks. A deployed detector must survive users who actively evade it, so Figure 5(a) probes eight evasion strategies, four perturbation and four paraphrasing, ordered by rewriting depth: character-level noise displaces a few surface tokens, whereas paraphrasing preserves meaning but rewrites the entire surface, directly negating a vocabulary-level signature. VisScore+LGBM is the most robust entry, averaging 87.5 AUC across the eight strategies and best on seven of them; the perturbation half costs it only 1.3 points from its clean mean, the paraphrasing half 20.9, while the scalar VisScore averages 76.2. The two baseline families fail differently: RepreGuard collapses on every axis to 54.5–63.3, its fitted projection erased by any surface rewriting, while XLM-RoBERTa averages 81.0 but falls from 95.0 to 33.5 on seq2seq paraphrasing attack, below chance because its learned direction reverses.

Varying lengths. Figure 5(c) probes the effect of test-text length, truncating DRX documents to 64–512 tokens. Shorter texts degrade every method, but by different margins: VisScore+LGBM is the strongest at every length, climbing from 86.1 AUC at 64 tokens back to 98.6 at 512; the scalar VisScore matches XLM-RoBERTa up to 128 tokens but trails on longer texts, 87.2 against 96.0 at 512; RepreGuard stays low at 58.6–62.8 across all four lengths.

6 Conclusion

We introduced VisScore, an LLM-generated text detector built on the concept-visibility hypothesis: a frozen reader's intermediate computation, decoded into vocabulary space, surfaces the stylistic signature of LLM-generated text as small, readable vocabularies. Scoring each document by the difference of LLM- and human-concept hit rates at a single golden layer, VisScore outperforms all 11 baselines on DetectRL-X, transfers to unseen languages, generators, and domains at a cost of at most 0.9 AUC points, and needs only a few labeled pairs to calibrate; a $\sim 10^4$-parameter head over the concept-hit statistics recovers the structure the scalar score discards. Since the decoded words are simultaneously the feature and the evidence, every verdict can be read and audited by a human reviewer, turning detection from an opaque score into an inspectable decision.

AI USE STATEMENT

During the preparation of this work, the authors used the coding assistant to assist with English language polishing and code debugging. All experiments were designed and executed by the authors, and every reported result was generated using the authors' own experimental pipeline. The authors reviewed and verified all AI-assisted outputs and take full responsibility for the content of this publication.

REFERENCES

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Bpcgcr8E8Z.

Advik Raj Basani and Pin-Yu Chen. Diversity boosts ai-generated text detection. *Trans. Mach. Learn. Res.*, 2026, 2026. URL https://openreview.net/forum?id=WscAU9It1l.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. volume abs/2303.08112, 2023. doi: 10.48550/ARXIV.2303.08112. URL https://doi.org/10.48550/arXiv.2303.08112.

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. *Trans. Assoc. Comput. Linguistics*, 13:1812–1831, 2025. doi: 10.1162/TACL.A.61. URL https://doi.org/10.1162/tacl.a.61.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.747. URL https://doi.org/10.18653/v1/2020.acl-main.747.

Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in education and teaching international*, 61(2):228–239, 2024.

Evan Crothers, Nathalie Japkowicz, and Herna L. Viktor. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002, 2023. doi: 10.1109/ACCESS.2023.3294090. URL https://doi.org/10.1109/ACCESS.2023.3294090.

Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Kaiyuan Zhang, Guanhong Tao, Guangyu Shen, and Xiangyu Zhang. Biscope: Ai-generated text detection by checking memorization of preceding tokens. 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/bc808cf2d2444b0abcceca366b771389-Abstract-Conference.html.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul C. Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *CoRR*, abs/2607.15495, 2026. doi: 10.48550/ARXIV.2607.15495. URL https://doi.org/10.48550/arXiv.2607.15495.

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 Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of

Proceedings of Machine Learning Research, pp. 17519–17537. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/hans24a.html.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27): eadt3813, 2025.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

nostalgebraist. Interpreting GPT: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020.

Artidoro Pagnoni, Martin Graciarena, and Yulia Tsvetkov. Threat scenarios and best practices to detect neural fake news. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pp. 1233–1249. International Committee on Computational Linguistics, 2022. URL https://aclanthology.org/2022.coling-1.106.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. Release strategies and the social impacts of language models, 2019. URL http://arxiv.org/abs/1908.09203.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pp. 12395–12412. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.827. URL https://doi.org/10.18653/v1/2023.findings-emnlp.827.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do llms exhibit human-like response biases? A case study in survey design. *Trans. Assoc. Comput. Linguistics*, 12:1011–1026, 2024. doi: 10.1162/TACL_A_00685. URL https://doi.org/10.1162/tacl_a_00685.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. Seqxgpt: Sentence-level ai-generated text detection. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 1144–1156. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.73. URL https://doi.org/10.18653/v1/2023.emnlp-main.73.

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. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b61bdf7e9f64c04ec75a26e781e2ad51-Abstract-Datasets_and_Benchmarks_Track.html.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia S. Chao, and Derek Fai Wong. A survey on llm-generated text detection: Necessity, methods, and future directions. *Comput. Linguistics*, 51(1):275–338, 2025. doi: 10.1162/COLI_A_00549. URL https://doi.org/10.1162/coli_a_00549.

Junchao Wu, Yefeng Liu, Chenyu Zhu, Hao Zhang, Zeyu Wu, Tianqi Shi, Yichao Du, Longyue Wang, Weihua Luo, Jinsong Su, and Derek F. Wong. Detectrl-x: Towards reliable multilingual and real-world llm-generated text detection. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 38247–38294. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1773. URL https://doi.org/10.18653/v1/2026.acl-long.1773.

A Concept Evidence

This appendix collects VisScore's evidence artifacts at the word level and at the span level. Table 4 contrasts the two signature vocabularies of §3.2 on DRX English, ranked by the per-token differential $\Delta(v)$ of Eq. 3: the LLM side is uniformly "flowery prose," whereas the human side is casual, functional vocabulary, with British spellings *programme* and *centre* reflecting the corpus. The LLM register is sharp, reaching $\Delta = +4.18$ at the top and $+1.87$ still at rank 100, while every human differential stays below 1.2 in magnitude: human writing samples a diffuse register rather than converging on one, echoing the asymmetric structure of Figure 1b. Non-Latin tokens are cross-lingual noise in DRX and are filtered from display, while alphabetic BPE fragments (e.g., *icions*, *TEMPL*) pass the readability filter and remain part of the concept sets; the full $500{+}325$ token sets are released with the code.

Figure 7 shows the span-granularity evidence visualization ($S{=}32$) discussed in §3.2, on a matched DetectRL test pair.

B Exact Definitions of the Visibility-Facet Features

This appendix lists the exact definitions of the 15 visibility-facet features summarized in §3.2. Let $a_i \in \{0, 1\}$ (resp. b_i) indicate that the workspace top-K at position i contains at least one LLM-signature (resp. human-signature) concept, and let $r \in \{0, 1\}^n$ denote either hit sequence; a *run* is a maximal block of consecutive hits. The following six statistics are computed for each of the two hit sequences separately:

- **Hit rate** (density): $\frac{1}{n}\sum_{i=1}^{n} r_i$.
- **Lag-1 autocorrelation** (clustering): $\mathrm{corr}(r_{1:n-1},\, r_{2:n})$.
- **Maximal run length** (clustering): the length of the longest run of consecutive hits, divided by n.
- **Burst fraction** (clustering): the share of hits that lie in runs of length at least two.
- **Center of mass** (position): $\frac{1}{n}\,\mathrm{mean}\{i : r_i = 1\}$, the mean position of the hits.
- **First-/second-half trend** (position): the share of hits in the second half of the text minus that in the first.

Three further exclusivity features compare the two sequences position by position:

- **LLM-only rate**: $\frac{1}{n}\sum_i a_i(1{-}b_i)$, the rate of positions that hit LLM-signature concepts only.
- **Human-only rate**: $\frac{1}{n}\sum_i (1{-}a_i)b_i$, the rate of positions that hit human-signature concepts only.
- **Joint-hit rate**: $\frac{1}{n}\sum_i a_i b_i$, the rate of positions that hit both.

Computing the six statistics for each side and adding the three rates yields the feature vector $\phi(\mathcal{T}) \in \mathbb{R}^{15}$.

Table 4: Top-100 signature concept tokens per side on DRX English, ranked by the per-token differential $\Delta(v)$ of Eq. 3 at the golden layer $l^\star$=22 from 9,600 train pairs; positive Δ favors LGT. Non-Latin tokens are cross-lingual noise in DRX and are filtered from display.

#	LLM concept	Δ	Human concept	Δ	#	LLM concept	Δ	Human concept	Δ
1	meticulously	+4.18	mainly	−1.15	51	showcases	+2.16	enlarg	−0.68
2	unparalleled	+3.49	Cumhurbaşkan	−1.15	52	showcase	+2.13	above	−0.67
3	vibrant	+3.38	relating	−1.08	53	encompass	+2.13	called	−0.67
4	effortlessly	+3.33	other	−0.94	54	unleash	+2.13	Related	−0.67
5	captivating	+3.33	sometime	−0.93	55	transformative	+2.12	mostly	−0.67
6	seamlessly	+3.19	also	−0.93	56	embody	+2.11	specially	−0.67
7	alongside	+3.10	Other	−0.91	57	intriguing	+2.11	InThe	−0.66
8	showcasing	+3.04	Also	−0.90	58	Navig	+2.10	complied	−0.66
9	adept	+2.96	used	−0.88	59	subtly	+2.09	appropriate	−0.66
10	firsthand	+2.95	programme	−0.86	60	culinary	+2.09	corresponding	−0.66
11	intricate	+2.84	very	−0.85	61	boasts	+2.07	arou	−0.66
12	advancements	+2.81	related	−0.84	62	ellar	+2.06	because	−0.66
13	amidst	+2.78	separately	−0.83	63	align	+2.06	Respond	−0.65
14	renowned	+2.76	etc	−0.83	64	undeniable	+2.05	attached	−0.65
15	meticulous	+2.75	pictured	−0.80	65	unprecedented	+2.05	really	−0.65
16	cherished	+2.58	centre	−0.80	66	bustling	+2.05	possibly	−0.65
17	compelling	+2.53	indicated	−0.78	67	Navigate	+2.05	behaviour	−0.65
18	antro	+2.52	administr	−0.78	68	aspiring	+2.04	emphas	−0.64
19	leveraging	+2.51	totally	−0.78	69	challenges	+2.04	interested	−0.64
20	thriving	+2.49	actually	−0.76	70	tirelessly	+2.04	organised	−0.64
21	showcased	+2.48	included	−0.76	71	thrilling	+2.03	Use	−0.64
22	delve	+2.47	probably	−0.76	72	rikes	+2.03	Internet	−0.64
23	countless	+2.46	therefore	−0.75	73	beneath	+2.02	behavioural	−0.64
24	envision	+2.45	according	−0.75	74	nestled	+2.02	accepted	−0.63
25	deeply	+2.40	presently	−0.75	75	societal	+2.01	signed	−0.63
26	fostering	+2.40	mentioned	−0.74	76	pivotal	+2.01	Cumhurbaşkanı	−0.63
27	adorned	+2.37	purposely	−0.74	77	icions	+2.00	eroon	−0.63
28	akin	+2.36	pertaining	−0.74	78	safeguard	+2.00	programmes	−0.63
29	engaging	+2.35	which	−0.72	79	reminiscent	+2.00	Express	−0.63
30	enduring	+2.35	produced	−0.72	80	ancements	+1.99	İnternet	−0.62
31	navigating	+2.32	started	−0.72	81	StateManager	+1.98	usually	−0.62
32	invaluable	+2.31	İnt	−0.72	82	impactful	+1.98	Included	−0.62
33	groundbreaking	+2.28	unsuccessfully	−0.71	83	transforming	+1.98	listed	−0.62
34	nurturing	+2.28	afterwards	−0.71	84	nurt	+1.97	Please	−0.61
35	intertwined	+2.27	apparently	−0.71	85	orchestrated	+1.97	Based	−0.61
36	crafted	+2.27	concerned	−0.71	86	intertw	+1.97	obtained	−0.61
37	unmatched	+2.26	stated	−0.71	87	relentless	+1.96	supplied	−0.61
38	undoubtedly	+2.25	jointly	−0.71	88	encaps	+1.96	installed	−0.60
39	burgeoning	+2.24	association	−0.71	89	crucial	+1.93	some	−0.60
40	foundational	+2.23	partly	−0.70	90	wield	+1.93	obviously	−0.60
41	tailored	+2.21	preferably	−0.70	91	steadfast	+1.93	practise	−0.60
42	embrace	+2.20	analys	−0.70	92	interconnected	+1.92	shown	−0.60
43	insights	+2.20	filed	−0.70	93	capt	+1.92	Images	−0.60
44	evolving	+2.18	referred	−0.69	94	genuinely	+1.91	involved	−0.60
45	astery	+2.18	dated	−0.68	95	unveiling	+1.91	electronically	−0.60
46	crafting	+2.18	organisation	−0.68	96	TEMPL	+1.91	unpublished	−0.60
47	newfound	+2.17	aset	−0.68	97	seemingly	+1.91	Used	−0.59
48	swiftly	+2.17	published	−0.68	98	profound	+1.91	described	−0.59
49	resilience	+2.16	recognised	−0.68	99	grappling	+1.89	either	−0.59
50	intric	+2.16	showed	−0.68	100	albeit	+1.87	Others	−0.59

C Baseline Details

Shared scorer. All statistics-based baselines share the Qwen3-8B scoring model with bf16 precision and a max length of 1,024 tokens; Binoculars and Fast-DetectGPT additionally use Qwen3-8B-Base as the observer model, which shares the vocabulary. Log-Likelihood, Entropy, and LogRank average the per-token log-probability, full-vocabulary entropy, and exact rank of the observed token under the

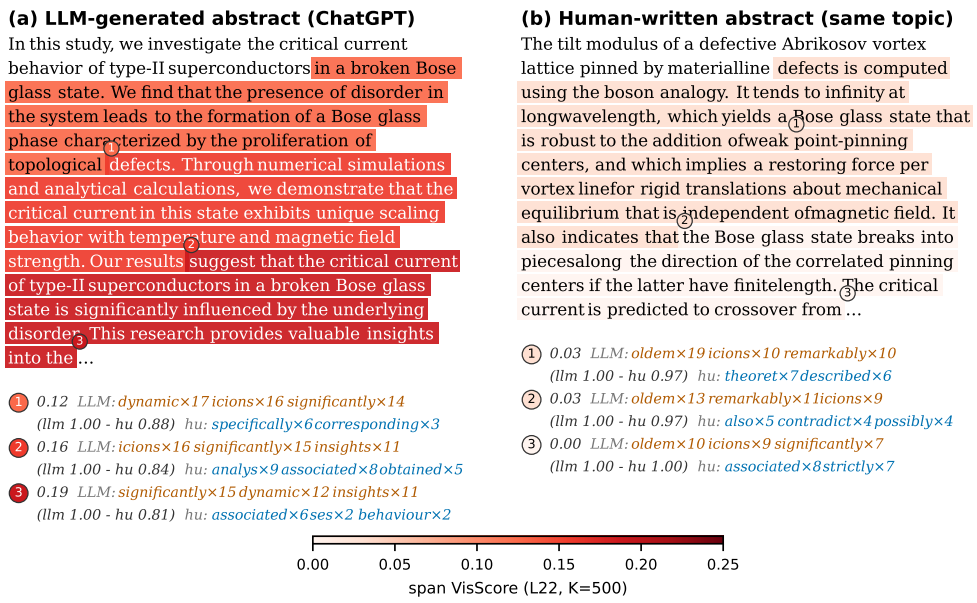


Figure 7: Workspace-concept evidence on a matched DetectRL test pair (Qwen3-8B reader, $l^\star{=}22$, $K{=}500$). Span intensity is proportional to its span VISSCORE (LLM-concept minus human-concept hit rate). Numbered notes list each span's most frequent concept hits (orange: LLM-signature, blue: human-signature). (a) LLM-generated abstract: the LLM-signature register saturates every span while the human-signature register lags (*significantly*, *dynamic*, *insights*), leaving a positive gap on every span. (b) Human-written abstract on the same topic: the two registers saturate together (*associated*, *corresponding*, *strictly*) and the gap all but vanishes: the discrimination lives in the difference, not in either register alone.

scorer. LRR averages the per-token ratio of log-probability to log-rank. Binoculars averages the gap between the observer's cross-entropy under the performer's token distribution and the observer's own token log-probability. Fast-DetectGPT draws one sample per position from the performer and scores the mean gap between the observer's log-probability of the observed token and of the sampled token.

RepreGuard. The surrogate model is Qwen3-8B. For each of its 36 layers, we fit a direction as the normalized difference between the mean last-token hidden states of the LLM-generated and human training documents, 9,600 per side per language, and RepreScore averages the token-level projections over positions and layers.

Training-based baselines. XLM-RoBERTa-base is fine-tuned per language for binary classification with learning rate 10^{-6}, 3 epochs, batch size 16, max length 512, weight decay 0.01, AdamW, bf16, and seed 2026. SeqXGPT keeps its official architecture and optimizer unchanged, a convolutional transformer with CRF trained by AdamW at learning rate 5×10^{-5} for 20 epochs with batch size 32 and sequence length 1,024; the four white-box scorer LLMs of the original are replaced by the single Qwen3-8B scoring model, whose token log-probability sequence is replicated across the four input channels. BiScope computes its official 72-dimensional features from Qwen3-8B, nine suffix segments crossed with the mean, max, min, and standard deviation of forward and backward cross-entropy, and keeps the official random-forest classifier with 100 trees. DivEye computes its official ten surprisal statistics from Qwen3-8B token surprisal, five distribution moments plus first- and second-order temporal differences, histogram entropy, and lag-1 autocorrelation, classified by the official XGBoost configuration with depth 12 and 200 trees.

Training-based head. The optional training-based head of §3.2 is a LightGBM classifier with row and feature subsampling rates of 0.9. Its remaining four hyperparameters are selected per language by 3-fold stratified cross-validation on the train split, maximizing the cross-validated AUC

over the grid n_estimators $\in \{300, 600\}$, learning rate $\in \{0.05, 0.1\}$, num_leaves $\in \{31, 63\}$, and min_child_samples $\in \{20, 50\}$. The selected configuration is refit on the full train split, and reported scores average the predicted probabilities over three seeds (11, 22, 33).

Compute. All experiments, including the reader forward passes of VISSCORE and all baselines, run on $4\times$ NVIDIA L40S GPUs with 48 GB memory each.

D TRANSFER FULL RESULTS: CROSS-LANGUAGE, CROSS-GENERATOR, AND CROSS-DOMAIN

This appendix collects the detailed transfer results summarized in Table 2 of Section 5.2: the full 8×8 train-on-one-language, test-on-all transfer matrix in Table 5, and the full 8-language leave-one-generator-out and leave-one-domain-out fold matrices for VISSCORE under the golden configuration in Tables 6 and 7. Within each fold, concepts are re-discovered on the train split of the remaining generators (or domains) and evaluated on the held-out generator's (domain's) test split, so the held-out factor never informs discovery. The Mean column macro-averages per-language AUCs; score scales differ across languages, which we quantify after the tables.

Table 5: Cross-language transfer matrix (AUC, golden configuration): train on the row language, evaluate on the column language. Diagonal (bold): in-language.

	Evaluation language								
Train	Ar	Zh	En	Fr	De	Pt	Ru	Es	Mean
Arabic	**98.4**	96.9	98.7	95.5	96.9	96.1	96.7	95.6	96.9
Chinese	96.3	**98.2**	98.5	94.9	94.5	95.1	94.5	96.6	96.1
English	97.6	98.8	**99.8**	98.7	97.9	98.3	96.4	98.7	98.3
French	96.6	97.5	99.4	**98.0**	96.9	97.4	94.6	98.4	97.4
German	96.6	96.6	98.7	96.4	**96.3**	95.1	93.9	96.7	96.3
Portuguese	96.8	97.1	99.6	98.1	97.0	**97.5**	95.3	98.2	97.4
Russian	94.9	95.7	96.0	93.4	93.0	92.2	**93.4**	92.7	93.9
Spanish	98.0	95.8	98.4	96.3	94.7	96.4	93.9	**97.5**	96.4

Table 6: Cross-generator generalization (AUC, golden configuration): leave-one-generator-out folds. Mean: macro-average over the eight languages.

Held-out	Language								
generator	Ar	Zh	En	Fr	De	Pt	Ru	Es	Mean
DeepSeek-V3	97.7	96.2	99.7	96.8	95.2	96.5	91.1	96.5	96.2
Gemini-2.5-Flash	98.9	99.4	99.9	98.4	97.3	98.8	95.5	98.2	98.3
GPT-4o	98.4	98.5	99.8	98.8	97.2	98.0	94.7	97.5	97.8
Qwen-Max	97.6	95.6	99.4	96.3	93.5	97.9	89.7	96.4	95.8

Table 7: Cross-domain generalization (AUC, golden configuration): leave-one-domain-out folds. Mean: macro-average over the eight languages.

Held-out	Language								
domain	Ar	Zh	En	Fr	De	Pt	Ru	Es	Mean
Academic	97.9	96.1	99.8	99.6	99.7	99.0	99.5	99.5	98.9
News	98.0	97.4	100.0	98.7	98.6	98.5	91.4	99.6	97.8
Novel	98.2	99.7	99.6	99.5	96.8	97.1	97.9	96.6	98.2
SEO	100.0	98.7	99.6	99.9	99.7	99.4	99.7	99.5	99.6
Webtext	97.3	97.5	99.9	99.3	97.2	99.1	93.9	99.6	98.0
Wiki	99.7	99.6	99.9	99.8	98.5	99.4	96.5	99.7	99.1

E LLM-GENERATION VARIANTS

Figure 5(b) probes the commonest in-the-wild pressure, which is not adversarial at all: users routinely condense, expand, or polish LLM-generated text before publishing. VISSCORE+LGBM averages 98.6 AUC across the three variants, exactly its clean-text mean, and expanding yields its best numbers of the entire evaluation, 99.6/96.4/96.9; VISSCORE averages 88.6, improving to 93.7 on expanding. XLM-RoBERTa's ranking is similarly robust at 96.0 mean AUC, but its thresholds fit on clean training text wobble, mean $F_1@1\%$ 56.3 against 93.8 for VISSCORE+LGBM; RepreGuard drops to 63.6, any LLM rewriting erasing its evidence. These strategies are attacks only in name: a further LLM pass injects more LLM-style vocabulary into the workspace, amplifying the signature rather than erasing it. Condensing is the only mild case for VISSCORE, 84.4, because compression deletes exactly the long, signature-dense spans.