DTS: Discriminant Truthfulness Scoring for Gradient-Free Hallucination Detection

Kaustubh Atey[1,*] **Shreeyam Bangera**[2,*] **Sandeep Kumar**[1] **Ankit Kumar**[1] **Amit Sethi**[1,3]

[1]Centre for Machine Intelligence and Data Science, IIT Bombay [2]Department of Physics, IIT Roorkee
[3]Department of Electrical Engineering, IIT Bombay *Equal contribution

EMNLP2026 BUDAPEST
GROUNDLM 2026 WORKSHOP
Grounding Language Models: Learning Faithfully and Efficiently
@ EMNLP 2026 · 29 October 2026

1 Problem & Motivation

- **Problem:** LLMs hallucinate as fluently and confidently as they answer correctly.
- **Scope:** Detection using only the generating model, with no retrieval, external judges, or extra models.
- **Existing approaches:** Confidence-based methods miss confident errors; sampling-based methods are costly and fail on repeated errors.
- **The gap:** Hidden-state detectors often use cosine or linear readouts that ignore anisotropic geometry.
- **Our question:** Can covariance-aware scoring improve detection without gradient training?

2 Core Insight & Key Findings

Core insight

- The **readout is a major bottleneck** in hidden-state detection, not just the steering strategy.
- **Covariance whitening** accounts for within-class variation, revealing signals cosine misses.

Key findings

- **Competitive accuracy:** DTS ranks first in 5 of 8 settings and second in the rest.
- **Better readout:** Whitening improves mean AUROC by **7.8 points** over cosine.
- **Efficient detection:** One forward pass and no gradient training: **1.0/1.2 min** vs. **10.2/12.3 min** for TSV⁺.

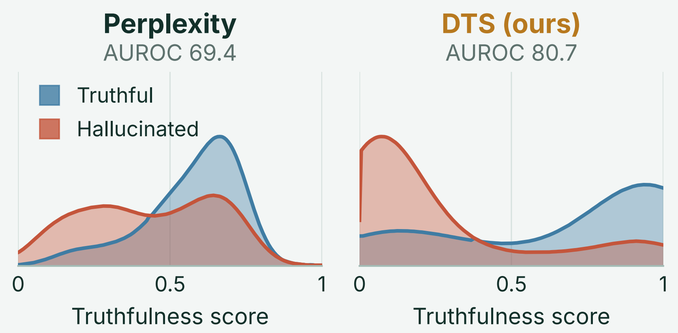
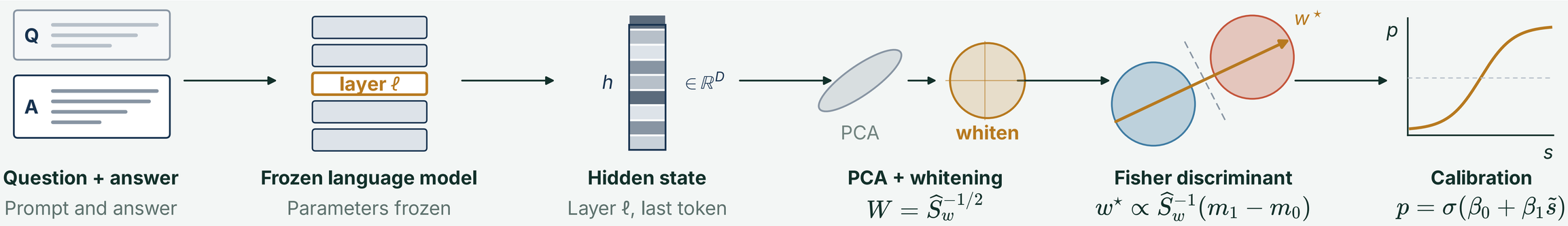


Figure. Llama-3.1-8B / TruthfulQA: DTS separates the classes that perplexity overlaps.

3 Methodology: DTS, Gradient-Free Truthfulness Scoring



Question + answer
Prompt and answer

Frozen language model
Parameters frozen

Hidden state
Layer ℓ, last token

PCA + whitening
$W = \widehat{S}_w^{-1/2}$

Fisher discriminant
$w^\star \propto \widehat{S}_w^{-1}(m_1 - m_0)$

Calibration
$p = \sigma(\beta_0 + \beta_1 \tilde{s})$

Representation Extraction

- **Frozen LLM:** Parameters are never updated; extract last-token hidden states at every layer.
- Select the most informative layer by 5-fold CV.

Closed-form Discriminant

- PCA projects h to $z \in \mathbb{R}^p$, $p = 200$.
- Whiten z by $W = \widehat{S}_w^{-1/2}$ (regularized).
- Fisher direction $\mathbf{d} \propto W(m_1 - m_0)$; no gradient descent.

Calibration

- Raw score: $s = \mathbf{d}^\top W z$
- Standardized score: $\tilde{s} = (s - \bar{s})/\mathrm{sd}(s)$
- Fit two logistic parameters $(\beta_0,\ \beta_1)$.

OPTIONAL Causal Steering

- Add the class-mean difference at one layer.
- Select layer and $\alpha \in \{0, 4, 8, 16\}$ by CV; $\alpha = 0$ is off.

$$\tilde{h} = h + \alpha(\mu_1 - \mu_0)$$

Frozen backbone • Closed-form discriminant • Two-parameter logistic calibration • No repeated generations required

4 Main Results

Method	Qwen-2.5-7B				Llama-3.1-8B			
	TQA	Triv	SciQ	NQ	TQA	Triv	SciQ	NQ
Perplexity	65.1	50.2	53.4	51.2	71.4	76.3	52.6	50.3
Sem. Entropy	66.1	58.7	65.9	65.3	59.4	68.7	68.2	60.7
HaloScope	81.3	73.4	76.6	65.7	70.6	76.2	76.1	62.7
SAPLMA†	81.7	82.0	81.5	67.9	78.2	83.7	77.3	62.8
TSV†	88.7	84.2	84.8	**76.2**	**85.5**	87.2	**88.6**	78.0
DTS (ours)	**90.1**	**91.6**	**86.2**	75.3	82.7	**90.9**	87.6	**78.9**

Table. Test AUROC (%). TQA: TruthfulQA; Triv: TriviaQA; NQ: NQ-Open. † Fully supervised; best in bold, second-best underlined.

- **DTS achieves the best AUROC in 5 of 8 settings** and ranks second to fully supervised TSV⁺ in the remaining three.
- Outperforms all other baselines, including supervised probes, hidden-state detectors, and sampling-based methods.

5 When Steering Helps

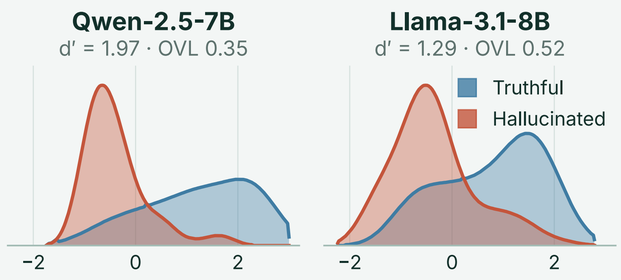


Figure. Held-out discriminant-score distributions on TruthfulQA: Qwen (left) and Llama (right).

- **Qwen:** Better separation (d′ = 1.97, OVL = 0.35); CV selects no steering on all four datasets.
- **Llama:** More overlap (d′ = 1.29, OVL = 0.52); steering lifts TruthfulQA AUROC from **80.7 to 82.7**.
- **Steering:** Mean-difference steering is comparable to trained TSV on Llama and generally beats random steering.

6 Cost versus Accuracy

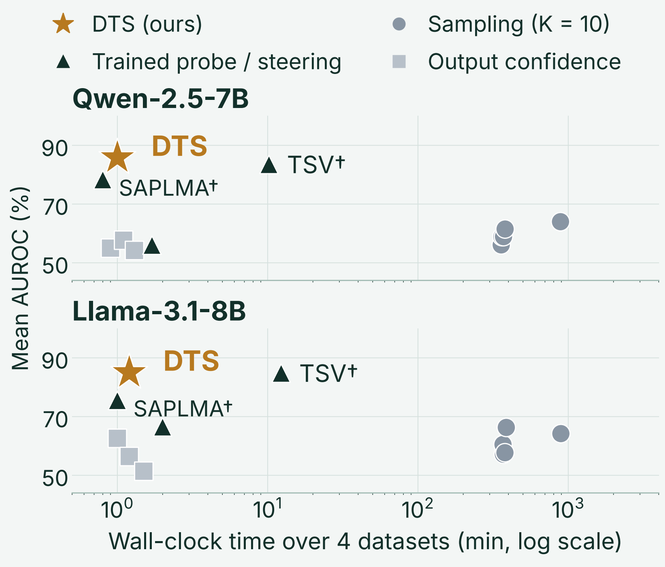


Figure. Mean AUROC vs. wall-clock time (4 datasets).

- **Accurate and efficient:** DTS reaches 85.8 / 85.0 AUROC, beating gradient-trained TSV⁺ (83.5 / 84.8).
- **~10× faster than TSV⁺:** 1.0 / 1.2 min vs. 10.2 / 12.3 min; one-time steering sweep totals 3.4 / 4.1 min.
- **Sampling is costly:** Multi-generation methods take hours, yet achieve at most 66.3 AUROC.

7 Long-form Generation

	DTS			Baselines	
Model	**Last**	**Mean**	**Max**	**PPL**	**LN-Ent**
Qwen-2.5-7B	74.8	88.0	**89.4**	71.0	69.3
Llama-3.1-8B	84.4	**89.5**	89.3	69.0	72.5

Table. Document-level AUROC (%) on FActScore-style biographies (n = 499 per model).

- **Task-dependent pooling:** Helps Llama but not Qwen on short QA (see box 8); helps both on long-form.
- **Strong gains:** Qwen 74.8 to 89.4 AUROC (max pooling); Llama 84.4 to 89.5 (mean).
- **Beats uncertainty baselines:** +18.4 over perplexity (Qwen), +17.0 over LN-Entropy (Llama); not span-level.

8 Representation Choice

- **Pooling helps on Llama:** mean AUROC 84.06 to **86.67/86.49**; last-token best on Qwen.
- **Llama/TruthfulQA:** Pooling reaches **85.1/86.0**, above steered last-token (82.7).
- **Per-token max:** 75.38 vs. 84.92 for last-token.

9 The Readout Is the Bottleneck

Readout	Qwen-2.5-7B				Llama-3.1-8B			
	TQA	Triv	SciQ	NQ	TQA	Triv	SciQ	NQ
Cosine	86.0	81.9	77.7	68.9	75.2	82.6	76.9	67.5
Difference of means	85.9	81.1	78.2	68.0	71.7	82.7	77.5	67.1
Diagonal whitening	89.5	91.5	86.1	**75.4**	80.6	90.2	**86.9**	78.1
DTS (full)	**90.1**	**91.6**	**86.2**	75.3	**80.7**	**90.6**	86.7	**78.2**

Table. Readout ablation (AUROC %). Best per column in bold, DTS second underlined.

- **Whitening drives the gain:** +7.8 mean AUROC over cosine (+4.1 to +10.7).
- **Metric matters:** difference-of-means ≈ cosine.
- **Covariance awareness:** Regularized LDA matches DTS (84.93); linear SVM trails (82.84).

10 Conclusion & Limitations

Key Takeaways

- **Covariance-aware scoring** drives the gain.
- **Selective steering:** needed only on Llama.
- **Efficient:** single forward pass, no gradient training.

Limitations

- Needs labeled data; transfer inconsistent.
- Mean transfer gaps **12.8 (Qwen) / 14.8 (Llama)** AUROC, incl. oracle-polarity-corrected.
- BLEURT label noise; single-layer steering.

Scan for the paper
Contact kaustubh.atey@iitb.ac.in
GroundLM Workshop @ EMNLP 2026 · 29 October 2026

References SAPLMA: Azaria & Mitchell, Findings of EMNLP 2023 · HaloScope: Du et al., NeurIPS 2024 · TSV: Park et al., ICML 2025 · Semantic Entropy: Kuhn et al., ICLR 2023 · Fisher, Ann. Eugenics 1936 · Lee et al., NeurIPS 2018 · TruthfulQA: Lin et al., ACL 2022 · TriviaQA: Joshi et al., ACL 2017 · SciQ: Welbl et al., W-NUT 2017 · NQ-Open: Kwiatkowski et al., TACL 2019 · FActScore: Min et al., EMNLP 2023 · BLEURT: Sellam et al., ACL 2020 · Qwen2.5: Yang et al., arXiv 2024 · Llama 3: Grattafiori et al., arXiv 2024. Acknowledgment We thank the TSV authors for releasing their generations, labels, and evaluation splits.