Token-Efficient Change Detection in LLM APIs

Clément Lalanne*

Joint work with: Timothée Chauvin*, Erwan Le Merrer, Jean-Michel Loubes, François Taïani, Gilles Tredan

ANITI cnrs Inria INSTITUT de MATHEMATIQUES de TOULOUSE

Context

Current needs for AI deployment:

- AI **alignment**
- AI **robustness**
- **Security** across the entire AI supply chain
- …

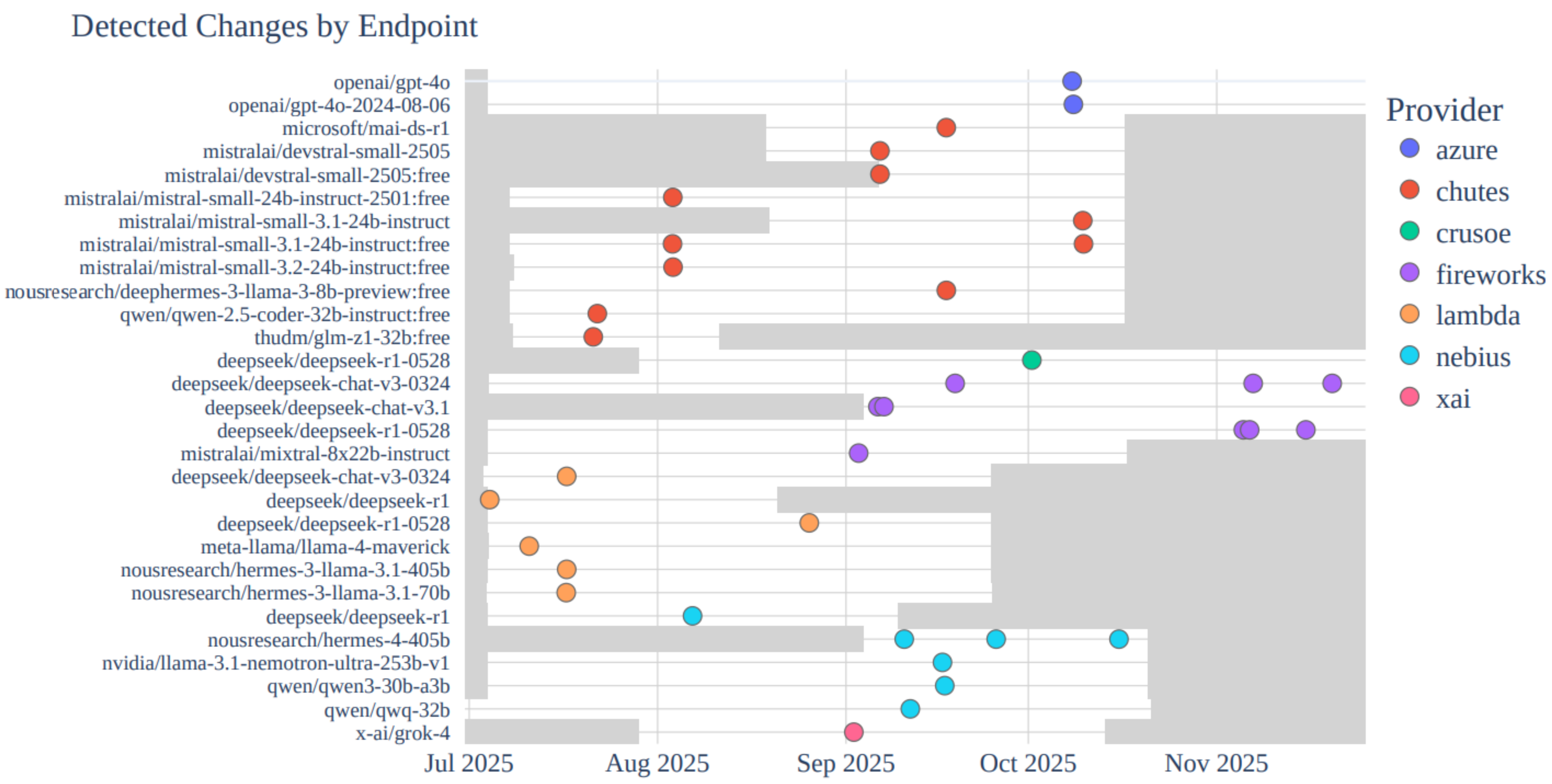
ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

LLM APIs are not stable

Reasons:

- 🆗 **Update** the model's behavior (should be disclosed)
- 👎 **Save costs** by quietly serving e.g. quantized model
- ‼️ Introduce **backdoors** (e.g. Malicious provider / supply chain attack / insider attack / bad update)
- …

Unexplained changes in production APIs



Source: Chauvin et al., Log Probability Tracking of LLM APIs, ICLR 2026

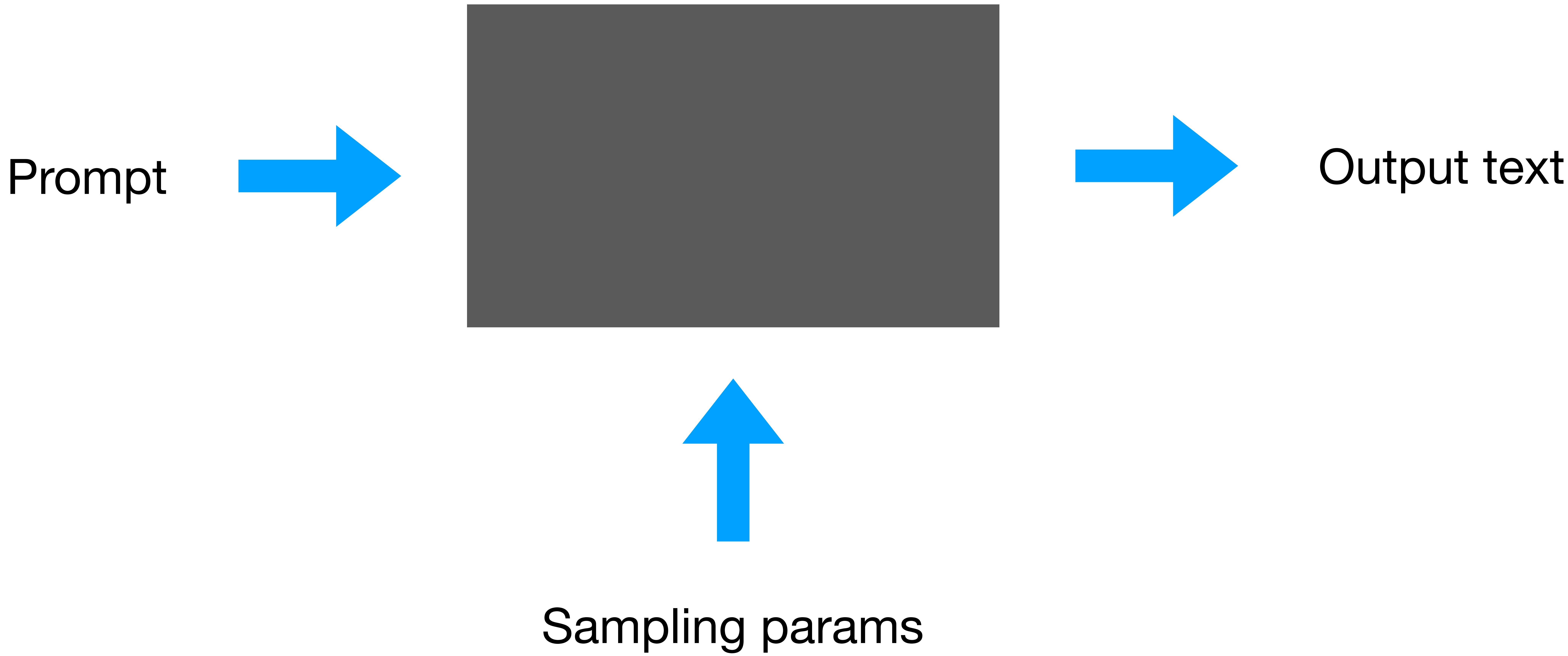
Goal

Monitor LLM APIs continuously to detect changes.

Requirements:

- **Cheap**
- Sensitive to **small changes**
- Works with **existing API providers**

Large Language Models 101



ANITI
INSTITUT de MATHEMATIQUES de TOULOUSE

Continue the sequence :
4
8
15
16

google/gemma-4-e4b

The pattern appears to be based on two separate, interwoven sequences (subsequences) that alternate every step.

1. **Subsequence A (Odd positions):** 4, 15, ...
   - $15 - 4 = 11$.
   - If we assume the difference is constant (+11), the next term in this subsequence would be $15 + 11 = 26$.
2. **Subsequence B (Even positions):** 8, 16, ...
   - $16 - 8 = 8$.
   - This sequence is increasing by 8. The next term in this subsequence would be $16 + 8 = 24$.

The next number requested is the fifth term (a_5), which belongs to Subsequence A.

Therefore, the continuation is **26**.

(Sequence continues: 4, 8, 15, 16, **26**, 24, ...)

Settings

Temperature	1
Limit Response Length	☐
Context Overflow	Truncate Middle
Stop Stri...	Enter a string and press ⏎
CPU Threads	7

Sampling

Top K Sampling	64
Repeat Penalty ☑	1.1
Presence Penalty ☐	
Top P Sampling ☑	0.95
Min P Sampling ☑	0.05

LMStudio

ANITI

INSTITUT de MATHEMATIQUES de TOULOUSE

Autoregressive Modeling and Training

Notations:

$\mathscr{V} = \{1, \ldots, d\}$ (vocabulary, size d), $w = (w_1, \ldots, w_L) \in \mathscr{V}^L, w_{<t} := (w_1, \ldots, w_{t-1})$

Text as a conditional probabilistic model:

$$p_\theta(w) = \prod_{t=1}^{L} p_\theta\left(w_t \mid w_{<t}\right)$$

ANITI

INSTITUT de MATHEMATIQUES de TOULOUSE

$\mathscr{V} = \{1,\ldots,d\}$ (vocabulary, size d), $w = (w_1, \ldots, w_L) \in \mathscr{V}^L$, $w_{<t} := (w_1, \ldots, w_{t-1})$

$$p_\theta(w) = \prod_{t=1}^{L} p_\theta\big(w_t \mid w_{<t}\big)$$

Architecture:

$$\underbrace{r(\theta_{\text{pre}}, w_{<t})}_{\in \mathbb{R}^m} \xrightarrow{\text{linear head } (W,b)} z(\theta, w_{<t}) = W\, r(\theta_{\text{pre}}, w_{<t}) + b \in \mathbb{R}^d, \qquad \theta = (\theta_{\text{pre}}, W, b)$$

Feature vector

Logprobs

$$p_\theta\big(w_t = i \mid w_{<t}\big) = f_i(\theta, w_{<t}) := \frac{\exp\big(z_i(\theta, w_{<t})/\tau\big)}{\sum_{j=1}^{d} \exp\big(z_j(\theta, w_{<t})/\underbrace{\tau}_{>0}\big)}$$

$\mathscr{V} = \{1,\ldots,d\}$ (vocabulary, size d), $w = (w_1,\ldots,w_L) \in \mathscr{V}^L, w_{<t} := (w_1,\ldots,w_{t-1})$

$$p_\theta(w) = \prod_{t=1}^{L} p_\theta(w_t \mid w_{<t})$$

$$\underbrace{r(\theta_{\text{pre}}, w_{<t})}_{\in \mathbb{R}^m} \xrightarrow{\text{linear head } (W,b)} z(\theta, w_{<t}) = W\,r(\theta_{\text{pre}}, w_{<t}) + b \in \mathbb{R}^d, \qquad \theta = (\theta_{\text{pre}}, W, b)$$

$$p_\theta(w_t = i \mid w_{<t}) = f_i(\theta, w_{<t}) = \frac{\exp\big(z_i(\theta, w_{<t})/\tau\big)}{\sum_{j=1}^{d} \exp\big(z_j(\theta, w_{<t})/\tau\big)}$$

Training:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \underbrace{-\mathbb{E}_{w\sim\mathscr{D}}\left[\sum_{t=1}^{L} \log p_\theta\big(w_t \mid w_{<t}\big)\right]}_{\mathscr{L}(\theta) \text{ (cross-entropy)}}$$

Empirical counterpart:

$$\hat{\mathscr{L}}(\theta) = -\frac{1}{|\mathscr{D}|}\sum_{w\in\mathscr{D}} \sum_{t=1}^{L} \log p_\theta\big(w_t \mid w_{<t}\big)$$

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

Inference

$$p_\theta\left(w_t = i \mid w_{<t}\right) \;=\; f_i(\theta, w_{<t}) \;=\; \frac{\exp\left(z_i(\theta, w_{<t})/\tau\right)}{\sum_{j=1}^{d} \exp\left(z_j(\theta, w_{<t})/\tau\right)}$$

Inference = sampling from the conditional, seeded by the prompt

θ fixed (deployed model), $x \in \mathscr{X}$ (prompt / context) $\qquad w_t \sim f(\theta, w_{<t}), \qquad t = 1,2,\ldots$

The temperature reshapes the head $\qquad M := \left\{ i : z_i(\theta, x) = \max_{1 \le j \le d} z_j(\theta, x) \right\}, \qquad k := |M|$

$\tau \to \infty : \; f \to \mathrm{Unif}(\mathscr{V}) \qquad \tau = 1 : f =$ trained distribution $\qquad \tau \to 0 : f \to \mathrm{Unif}(M)$

The prompt chooses where you probe $\qquad x \longmapsto z(\theta, x) \longmapsto M(x),\, k(x)$

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

Problem statement

API exposes:

$f(\theta, \cdot\,)$

User submits: (x, τ), observes $y = (y_1, \ldots, y_L)$, $\quad y_t \sim f\big(\theta, (x, y_{<t})\big)$.

The question: did the model change?

Query the model repeatedly at two time points: have its weights changed in between?

$H_0 : \theta_0 = \theta_1$ vs $H_1 : \theta_0 \neq \theta_1$

ANITI

INSTITUT de MATHEMATIQUES de TOULOUSE

First baseline: Model Equality Testing (MET)

Testing datasets:

$$\mathscr{D}_P = \Big\{ \big(x^{(i)}, y^{(i)}\big) \;:\; \underbrace{x^{(i)} \sim \pi, \quad y_t^{(i)} \sim f_{\theta_0}^{(\tau)}\big(\cdot \mid x^{(i)}, y_{<t}^{(i)}\big) \; \forall t}_{(x^{(i)}, y^{(i)}) \sim P := \pi f_{\theta_0}^{(\tau)}} \Big\}_{i=1}^{N} \qquad \mathscr{D}_Q = \Big\{ \big(x^{(i)}, y^{(i)}\big) \;:\; \underbrace{x^{(i)} \sim \pi, \quad y_t^{(i)} \sim f_{\theta_1}^{(\tau)}\big(\cdot \mid x^{(i)}, y_{<t}^{(i)}\big) \; \forall t}_{(x^{(i)}, y^{(i)}) \sim Q := \pi f_{\theta_1}^{(\tau)}} \Big\}_{i=1}^{N}$$

MMD score:

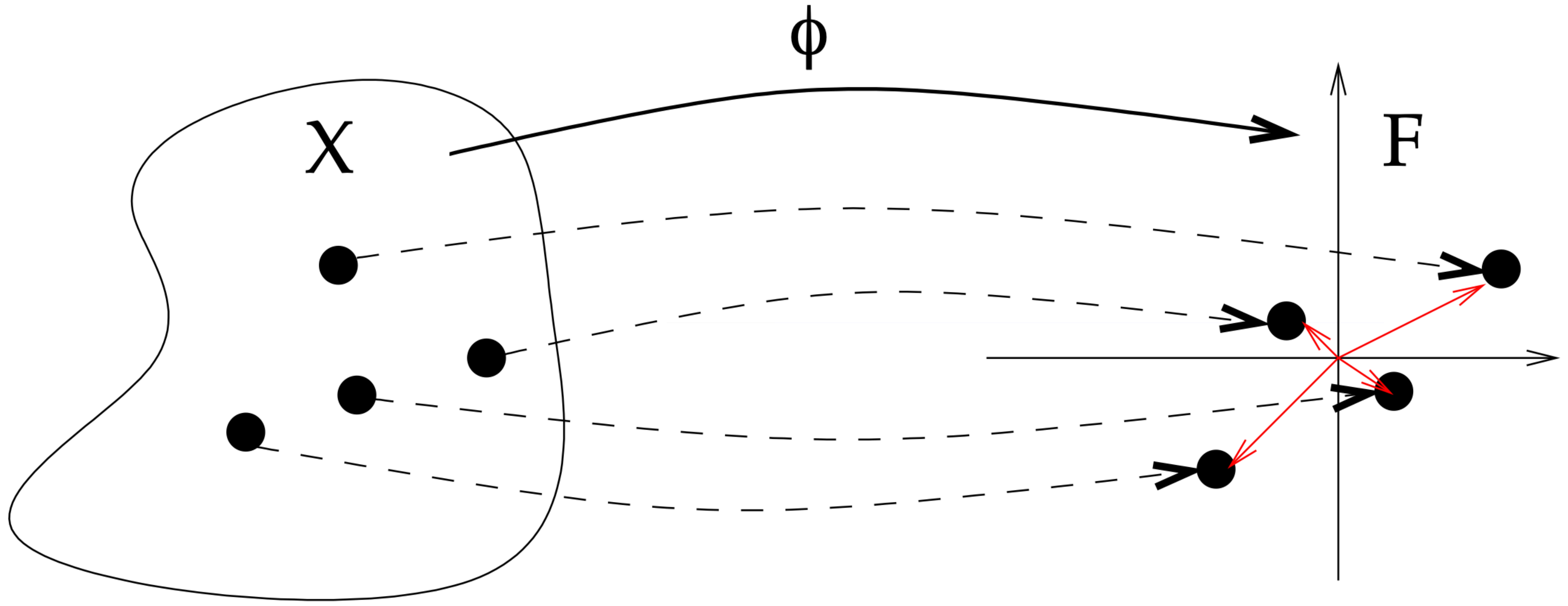
$$\mathrm{MMD}_\phi(P, Q) = \Big\| \, \mathbb{E}_{z \sim P}[\phi(z)] \; - \; \mathbb{E}_{z' \sim Q}[\phi(z')] \, \Big\|^2$$

$$\theta_0 = \theta_1 \implies P = Q \implies \mathrm{MMD}_\phi(P, Q) = 0 \implies \widehat{\mathrm{MMD}}(\mathscr{D}_P, \mathscr{D}_Q) \approx 0.$$

Source: Gao et al., Model Equality Testing, ICLR 2025

$$\mathscr{D}_P = \Big\{ \big(x^{(i)}, y^{(i)}\big) \ : \ \underbrace{x^{(i)} \sim \pi, \quad y_t^{(i)} \sim f_{\theta_0}^{(\tau)}\big(\,\cdot \mid x^{(i)}, y_{<t}^{(i)}\big) \ \forall t}_{(x^{(i)}, y^{(i)}) \,\sim\, P \,:=\, \pi f_{\theta_0}^{(\tau)}} \Big\}_{i=1}^{N} \quad \mathscr{D}_Q = \Big\{ \big(x^{(i)}, y^{(i)}\big) \ : \ \underbrace{x^{(i)} \sim \pi, \quad y_t^{(i)} \sim f_{\theta_1}^{(\tau)}\big(\,\cdot \mid x^{(i)}, y_{<t}^{(i)}\big) \ \forall t}_{(x^{(i)}, y^{(i)}) \,\sim\, Q \,:=\, \pi f_{\theta_1}^{(\tau)}} \Big\}_{i=1}^{N}$$

$$\mathrm{MMD}_{\phi}(P, Q) = \Big\| \, \mathbb{E}_{z \sim P}[\phi(z)] \ - \ \mathbb{E}_{z' \sim Q}[\phi(z')] \, \Big\|^2$$



Kernel trick: $\mathrm{MMD}_k(P, Q) = \mathbb{E}_{z,z' \sim P}[k(z, z')] + \mathbb{E}_{z,z' \sim Q}[k(z, z')] - 2\, \mathbb{E}_{\substack{z \sim P \\ z' \sim Q}}[k(z, z')]$

$$\widehat{\mathrm{MMD}}(\mathscr{D}_P, \mathscr{D}_Q) = \frac{1}{N(N-1)} \sum_{i \neq i'} k\Big((x^{(i)}, y^{(i)}), (x^{(i')}, y^{(i')})\Big) + \frac{1}{N(N-1)} \sum_{j \neq j'} k\Big((x^{(j)}, y^{(j)}), (x^{(j')}, y^{(j')})\Big) - \frac{2}{N^2} \sum_{i,j} k\Big((x^{(i)}, y^{(i)}), (x^{(j)}, y^{(j)})\Big)$$

Source: Julien Mairal's lecture on kernel methods

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

$$\mathcal{D}_P = \Big\{ \, (x^{(i)}, y^{(i)}) \ : \ \underbrace{x^{(i)} \sim \pi, \quad y_t^{(i)} \sim f_{\theta_0}^{(\tau)}(\,\cdot \mid x^{(i)}, y_{<t}^{(i)}) \ \forall t}_{(x^{(i)}, y^{(i)}) \, \sim \, P \, := \, \pi f_{\theta_0}^{(\tau)}} \, \Big\}_{i=1}^{N} \qquad \mathcal{D}_Q = \Big\{ \, (x^{(i)}, y^{(i)}) \ : \ \underbrace{x^{(i)} \sim \pi, \quad y_t^{(i)} \sim f_{\theta_1}^{(\tau)}(\,\cdot \mid x^{(i)}, y_{<t}^{(i)}) \ \forall t}_{(x^{(i)}, y^{(i)}) \, \sim \, Q \, := \, \pi f_{\theta_1}^{(\tau)}} \, \Big\}_{i=1}^{N}$$

$$\widehat{\mathrm{MMD}}(\mathcal{D}_P, \mathcal{D}_Q) = \frac{1}{N(N-1)} \sum_{i \neq i'} k\big((x^{(i)}, y^{(i)}), (x^{(i')}, y^{(i')})\big) + \frac{1}{N(N-1)} \sum_{j \neq j'} k\big((x^{(j)}, y^{(j)}), (x^{(j')}, y^{(j')})\big) - \frac{2}{N^2} \sum_{i,j} k\big((x^{(i)}, y^{(i)}), (x^{(j)}, y^{(j)})\big)$$

Exact kernel:

$$k\big((x, y), (x', y')\big) = \mathbf{1}\{x = x'\} \, \tilde{k}_{\text{hamming}}(y, y'), \qquad \tilde{k}_{\text{hamming}}(y, y') = \sum_{t=1}^{L} \mathbf{1}\{y_t = y'_t\}$$

Prompt distribution: Selected sentences from Wikipedia

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

✅ **Advantages:**

- Interpretable test statistic
- Interpretable failure mode
- Theoretically tractable in principle

❌ **Inconvenients:**

- Hamming kernel only sees the marginals
- Does not see what happens outside of the Wikipedia benchmark
- Quite expensive in practice ($67/year/model - hourly monitoring)

Second baseline: MMLU-ALG

Idea: monitor a fixed benchmark, watch the answer distribution drift.

Fixed prompt set: $\pi := \mathrm{Unif}\{x_1, \ldots, x_{P_{\mathrm{MA}}}\}, \qquad P_{\mathrm{MA}} = 100$

$\{x_1, \ldots, x_{100}\}$ = MMLU abstract_algebra subset (multiple-choice)

$$y^{(i)} \sim f(\theta, x_i), \qquad y^{(i)} \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$$

Why algebra? It is a difficult benchmark where models are usually uncertain.

✅ **Advantages:**

- Human-readable
- Rather cheap
- Theoretically tractable (with basic Hoeffding & Union bounds)

❌ **Inconvenients:**

- Low sensitivity to small changes
- Does not see what happens outside of the algebra benchmark

Our approach: Simplify the observation, then analyze the difficulty rigorously.

Fix $x_{\text{test}} \in \mathcal{X}$. $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} f(\theta, x_{\text{test}}) \in \mathcal{V}$ (n queries $\rightarrow$ n first tokens)

$\mathbf{p}_0 := f(\theta_0, x_{\text{test}})_{1:(d-1)}$, $\mathbf{p}_1 := f(\theta_1, x_{\text{test}})_{1:(d-1)}$

=> **The problem is a problem of multinomial testing !**

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

Fix $x_{\text{test}} \in \mathcal{X}$. $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} f(\theta, x_{\text{test}}) \in \mathcal{V}$ (n queries → n first tokens)

$\mathbf{p}_0 := f(\theta_0, x_{\text{test}})_{1:(d-1)}$, $\mathbf{p}_1 := f(\theta_1, x_{\text{test}})_{1:(d-1)}$

- Balakrishnan and Wasserman, Hypothesis testing for densities and high-dimensional multinomials, Annals of Stats., 2019
- Cai et al., Testing high-dimensional multinomials with applications to text analysis. JRSS-B, 2024
- Aliakbarpour et al., Optimal algorithms for augmented testing of discrete distributions, NeurIPS, 2024.
- Berrett et al. Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms, NeurIPS, 2020
- …

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

Fix $x_{\text{test}} \in \mathcal{X}$. $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} f(\theta, x_{\text{test}}) \in \mathcal{V}$ (n queries $\to$ n first tokens)

$\mathbf{p}_0 := f(\theta_0, x_{\text{test}})_{1:(d-1)}$, $\mathbf{p}_1 := f(\theta_1, x_{\text{test}})_{1:(d-1)}$

⚠️ Classical bounds depend on $(d, n, \|\mathbf{p}_0 - \mathbf{p}_1\|)$ only.

They never reveal how x_{test} and τ shape the difficulty

$d = |\mathcal{V}| \sim 10^5$, $n \ll d \;\Rightarrow$ most tokens never sampled

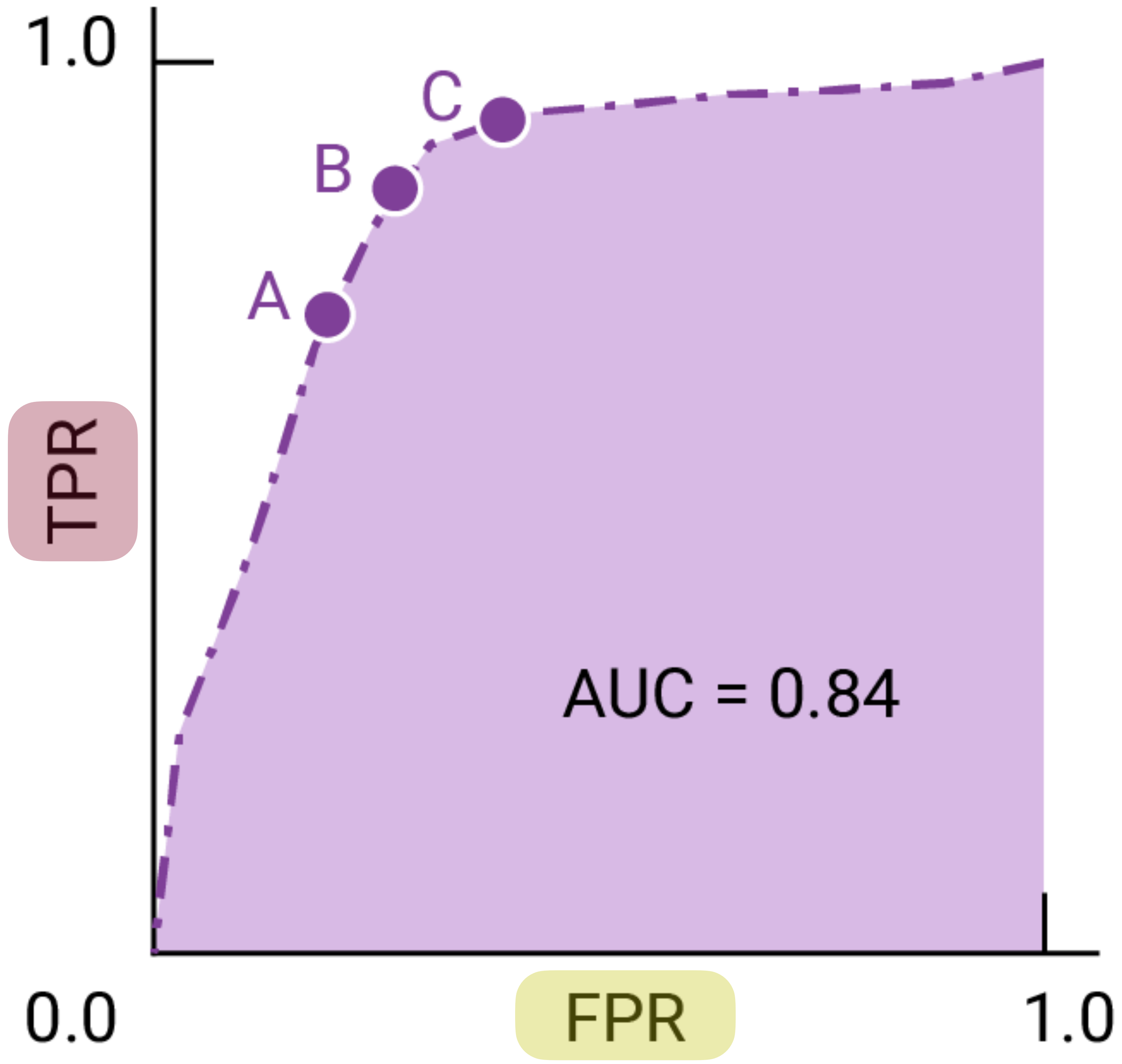
Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

Statistical tests 101

A test is a decision rule ϕ : (observations) $\to [0,1]$. $\phi = c$: reject H_0 with probability c .

$\alpha := \mathbb{E}_{H_0}[\phi] = \mathbb{E}[\phi \mid \theta_0 = \theta_1]$ (Type-I, false positive, FPR)

$\beta := \mathbb{E}_{H_1}[1-\phi] = 1 - \mathbb{E}_{H_1}[\phi]$ (Type-II, false negative, 1-TPR)



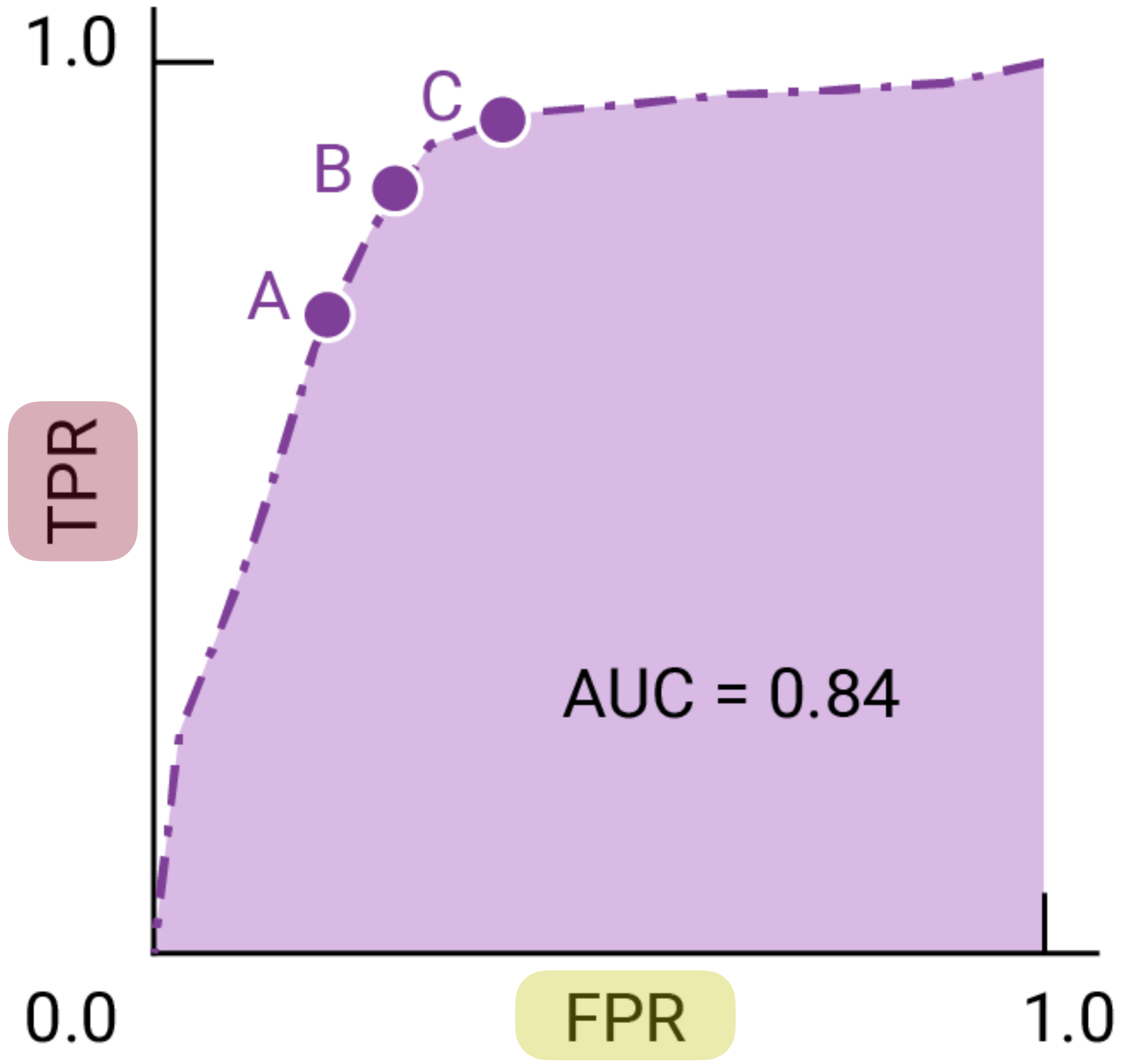
Source: Google for Dev.

A test is a decision rule ϕ : (observations) $\to [0,1]$. $\phi = c$: reject H_0 with probability c .

Trivial tests: $\phi \equiv 0 \Rightarrow \alpha = 0, \beta = 1.$ $\phi \equiv 1 \Rightarrow \alpha = 1, \beta = 0.$

A test is only defined by the $(\alpha, 1 - \beta)$ **pair**

Compare tests by their whole curve (or AUC), never by α or β alone.



Source: Google for Dev.

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

A test is a decision rule ϕ : (observations) $\to [0,1]$. $\phi = p$: reject H_0 with probability p.

Neyman-Pearson (informal). To test $H_0 : \mathbf{p}_0$ vs $H_1 : \mathbf{p}_1$,

among all tests with Type-I error $\leq \alpha$, the most powerful one

thresholds the **likelihood ratio** :

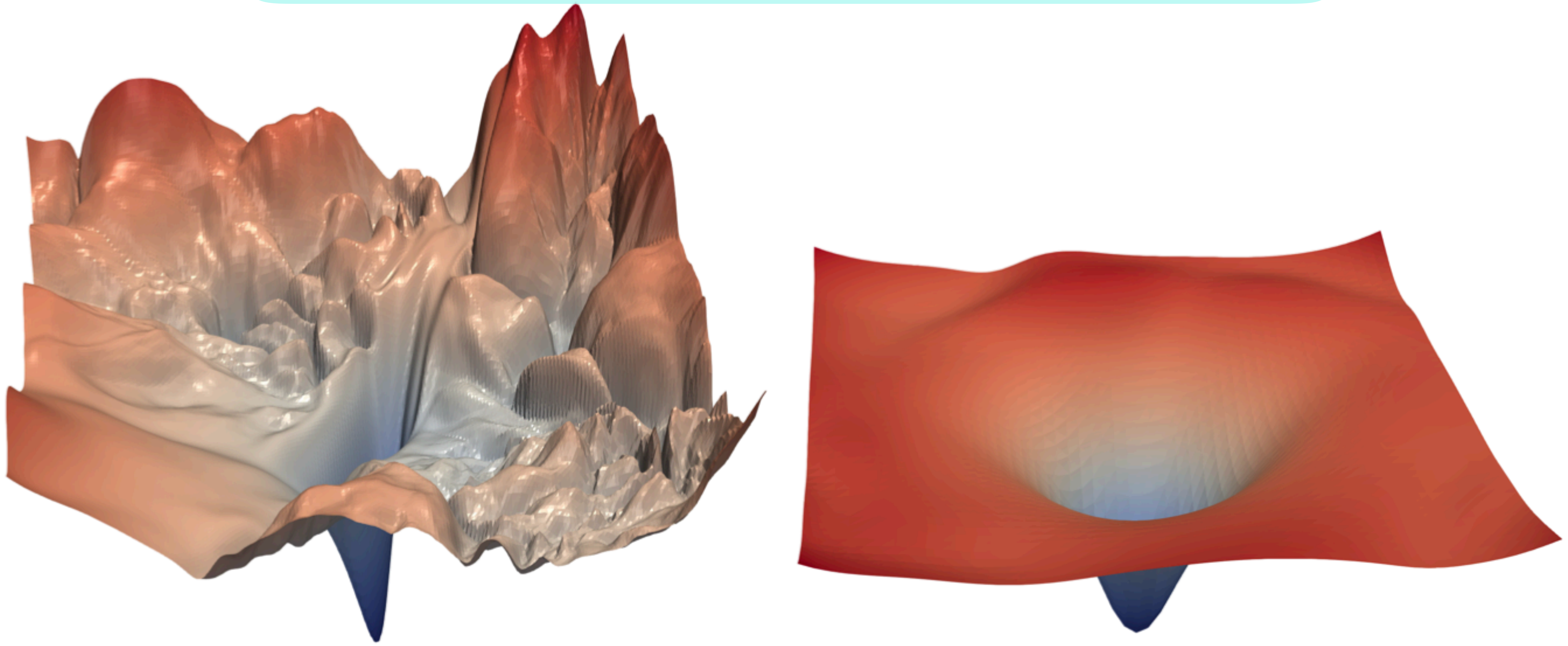
$$\Lambda(Y) = \frac{p_1(Y)}{p_0(Y)} \underset{H_0}{\overset{H_1}{\gtrless}} c\,.$$

The difficulty of eptimal tests is governed by how Λ **separates** $\mathbf{p}_0, \mathbf{p}_1$.

Generally, optimal tests are intractable

Fix $x_{\text{test}} \in \mathcal{X}$. $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} f(\theta, x_{\text{test}}) \in \mathcal{V}$ (n queries $\rightarrow$ n first tokens)

$$\log \Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_1, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}}.$$



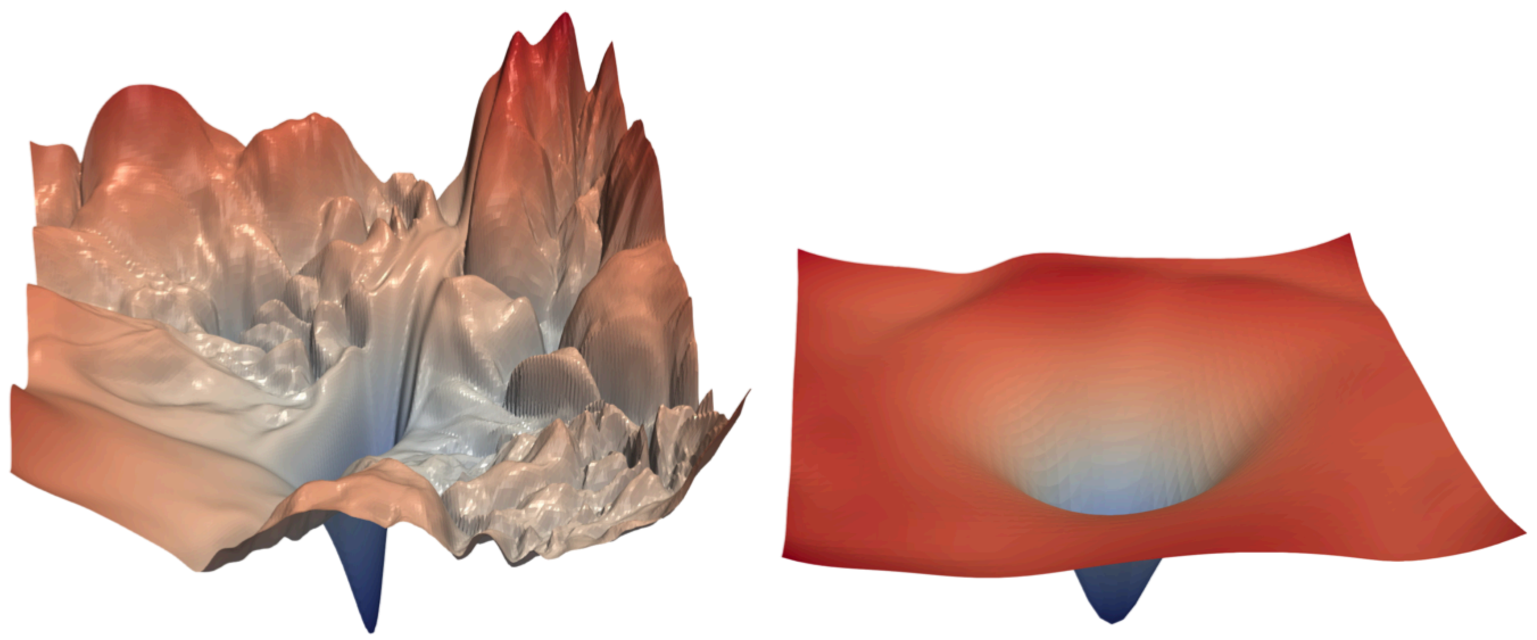
(a) without skip connections   (b) with skip connections

Source: Li et al., Visualizing the Loss Landscape of Neural Nets, Neurips, 2018

ANITI   INSTITUT de MATHEMATIQUES de TOULOUSE

Solution: we only look locally

$$\log \Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_1, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}}.$$



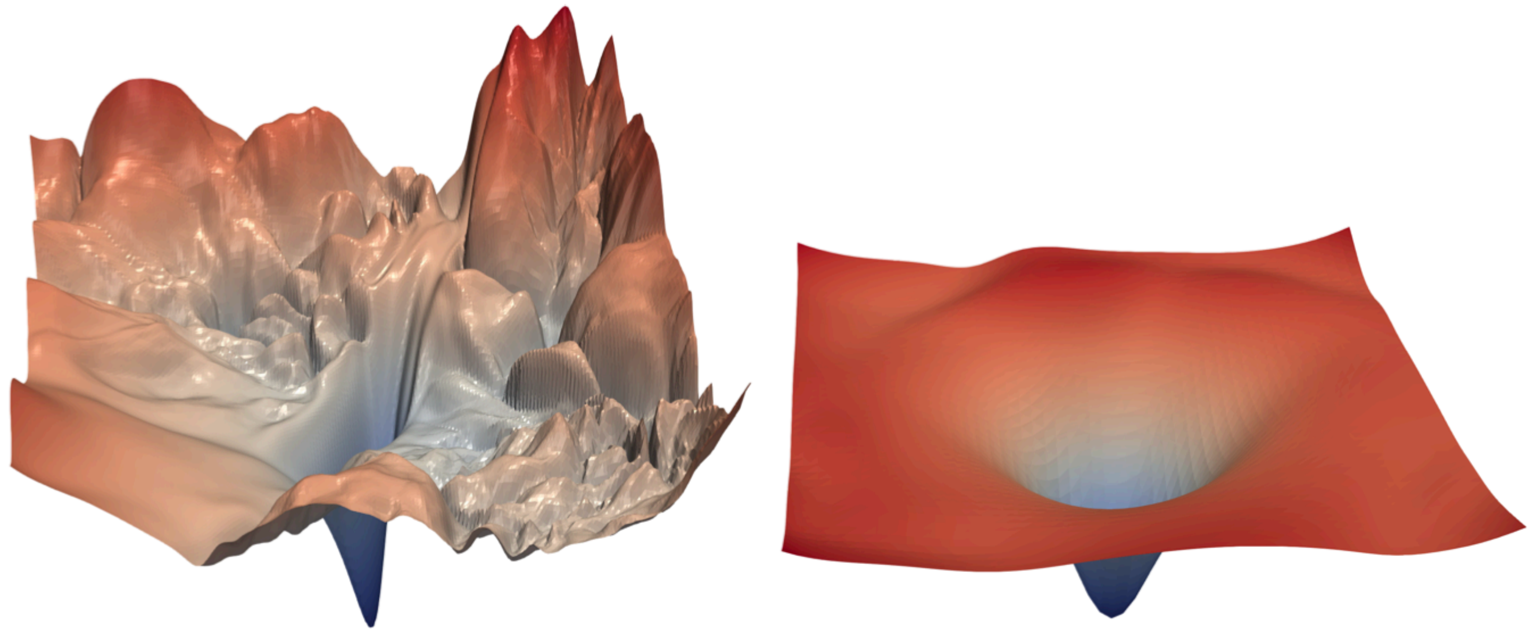
(a) without skip connections (b) with skip connections

Local Asymptotic Normality:

$$\epsilon_n = \frac{s}{\sqrt{n}}, \quad s \in \mathbb{R} \text{ fixed}, \mathbf{h} \text{ direction fixed}, \theta_1 = \theta_0 + \epsilon_n \mathbf{h}$$

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

$$\log \Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_1, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}}.$$



(a) without skip connections (b) with skip connections

$$\log \Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_0 + \epsilon_n \mathbf{h}, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}} = s\, \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{Z}_n - \frac{s^2}{2} \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{d} + o_{\mathbb{P}}(1),$$

$$\mathbf{d} := \nabla_\theta f(\theta_0, x_{\text{test}})_{1:(d-1)} \mathbf{h} \in \mathbb{R}^{(d-1)},$$

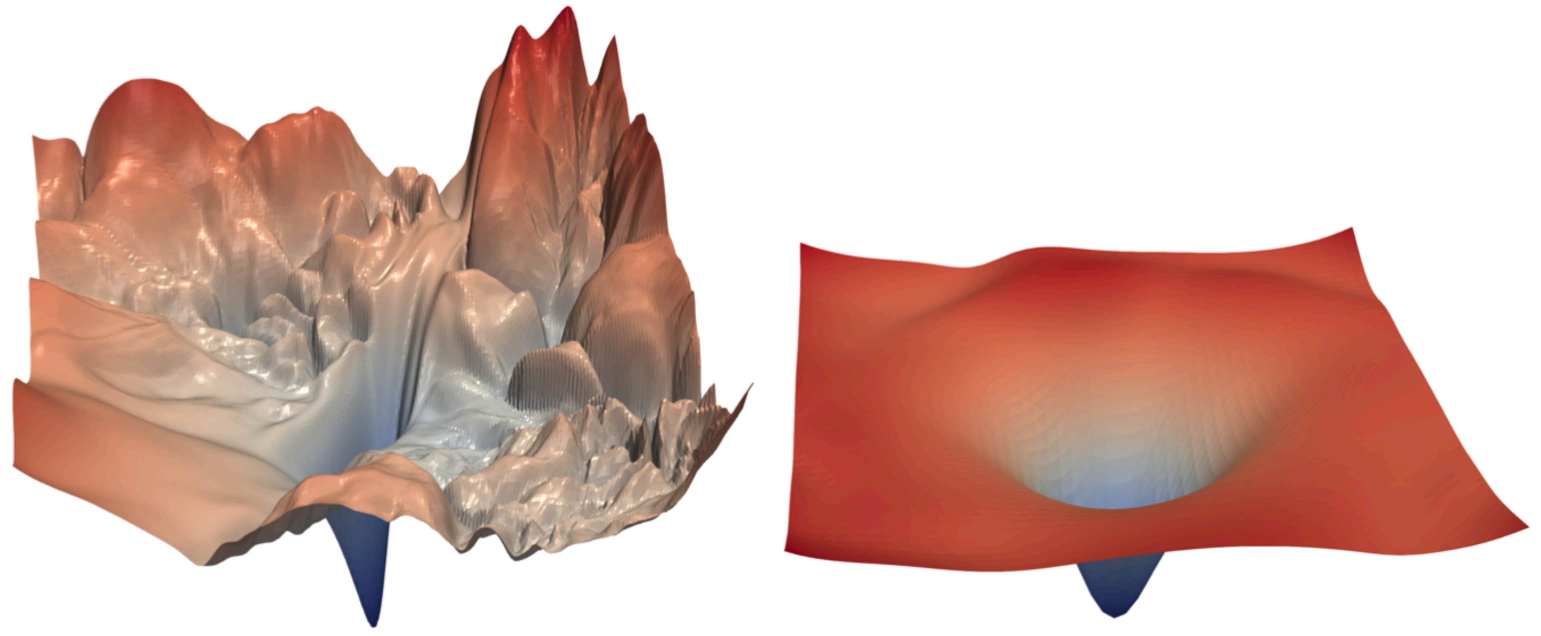
$$\mathbf{Z}_n := \sqrt{n}\,(\text{empirical token frequencies} - \mathbf{p}_0)\,.$$

$$F(\mathbf{p}) := \operatorname{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$$

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

$$\log \Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_1, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}} .$$



(a) without skip connections (b) with skip connections

$$\log \Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_0 + \epsilon_n \mathbf{h}, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}} = s\, \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{Z}_n - \frac{s^2}{2} \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{d} + o_{\mathbb{P}}(1),$$

$$\mathbf{d} := \nabla_\theta f(\theta_0, x_{\text{test}})_{1:(d-1)} \mathbf{h} \in \mathbb{R}^{(d-1)},$$

$$\mathbf{Z}_n := \sqrt{n}\,(\text{empirical token frequencies} - \mathbf{p}_0)\,.$$

$$F(\mathbf{p}) := \operatorname{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$$

Proof:

LLN under H0

Hoeffding under H1

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

$$\log \Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_0 + \epsilon_n \mathbf{h}, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}} = s\, \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{Z}_n - \frac{s^2}{2} \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{d} + o_{\mathbb{P}}(1),$$

$$\mathbf{d} := \nabla_\theta f(\theta_0, x_{\text{test}})_{1:(d-1)} \mathbf{h} \in \mathbb{R}^{(d-1)},$$

$$\mathbf{Z}_n := \sqrt{n}\,(\text{empirical token frequencies} - \mathbf{p}_0)\,.$$

$$F(\mathbf{p}) := \operatorname{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$$

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

$$\log \Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_0 + \epsilon_n \mathbf{h}, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}} = s\, \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{Z}_n - \frac{s^2}{2} \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{d} + o_{\mathbb{P}}(1),$$

$$\mathbf{d} := \nabla_\theta f(\theta_0, x_{\text{test}})_{1:(d-1)} \mathbf{h} \in \mathbb{R}^{(d-1)}$$

$\mathbf{Z}_n := \sqrt{n}$ (empirical token frequencies – $\mathbf{p}_0$) .

$$F(\mathbf{p}) := \operatorname{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$$

Under $H_0\ (\theta = \theta_0)$: $\quad \log \Lambda \xrightarrow{d} \mathcal{N}\left(-\frac{s^2}{2} \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{d},\ s^2 \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{d}\right),$

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

$$\log \Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_0 + \epsilon_n \mathbf{h}, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}} = s\, \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{Z}_n - \frac{s^2}{2} \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{d} + o_{\mathbb{P}}(1),$$

$$\mathbf{d} := \nabla_\theta f(\theta_0, x_{\text{test}})_{1:(d-1)} \mathbf{h} \in \mathbb{R}^{(d-1)}$$

$\mathbf{Z}_n := \sqrt{n}$ (empirical token frequencies $- \mathbf{p}_0$) .

$$F(\mathbf{p}) := \operatorname{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$$

Under $H_0\ (\theta = \theta_0)$: $\log \Lambda \xrightarrow{d} \mathcal{N}\left(-\frac{s^2}{2} \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{d},\ s^2 \mathbf{d}^\top F(\mathbf{p}_0)^{-1} \mathbf{d}\right),$

Proof: CLT

$$\log\Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_0 + \epsilon_n \mathbf{h}, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}} = s\,\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{Z}_n - \frac{s^2}{2}\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{d} + o_{\mathbb{P}}(1),$$

$\mathbf{d} := \nabla_\theta f(\theta_0, x_{\text{test}})_{1:(d-1)}\mathbf{h} \in \mathbb{R}^{(d-1)}$

$\mathbf{Z}_n := \sqrt{n}$ (empirical token frequencies – $\mathbf{p}_0$) .

$F(\mathbf{p}) := \mathrm{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$

Under $H_0\ (\theta = \theta_0)$: $\quad \log\Lambda \xrightarrow{d} \mathcal{N}\left(-\frac{s^2}{2}\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{d},\ s^2\,\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{d}\right),$

Proof: CLT

Under $H_1\ (\theta = \theta_0 + \epsilon_n \mathbf{h})$: $\quad \log\Lambda \xrightarrow{d} \mathcal{N}\left(+\frac{s^2}{2}\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{d},\ s^2\,\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{d}\right),$

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

$$\log \Lambda = \sum_{i=1}^{n} \log \frac{\left[f(\theta_0 + \epsilon_n \mathbf{h}, x_{\text{test}})\right]_{Y_i}}{\left[f(\theta_0, x_{\text{test}})\right]_{Y_i}} = s\,\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{Z}_n - \frac{s^2}{2}\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{d} + o_{\mathbb{P}}(1),$$

$\mathbf{d} := \nabla_\theta f(\theta_0, x_{\text{test}})_{1:(d-1)}\mathbf{h} \in \mathbb{R}^{(d-1)}$

$\mathbf{Z}_n := \sqrt{n}$ (empirical token frequencies $- \mathbf{p}_0$) .

$F(\mathbf{p}) := \mathrm{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$

Under $H_0\ (\theta = \theta_0)$: $\log \Lambda \xrightarrow{d} \mathcal{N}\left(-\frac{s^2}{2}\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{d},\ s^2\,\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{d}\right),$

Proof: CLT

Under $H_1\ (\theta = \theta_0 + \epsilon_n \mathbf{h})$: $\log \Lambda \xrightarrow{d} \mathcal{N}\left(+\frac{s^2}{2}\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{d},\ s^2\,\mathbf{d}^\top F(\mathbf{p}_0)^{-1}\mathbf{d}\right),$

Proof: Lévy to show the weak convergence of the pushforward measures

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

INSTITUT de MATHEMATIQUES de TOULOUSE

Theorem 3.1 (Optimal Tests in the LAN Regime). *Let $\alpha \in (0,1)$. If (ϕ_n) is a sequence of tests such that, for every n, ϕ_n is the test with the lowest Type-II error among all tests with Type-I error at most α in testing $\mathbf{p}_0$ vs $\mathbf{p}_n$, then*

$$\textit{Type-II}(\phi_n) \to \mathbb{P}\left(\mathcal{N}(0,1) \leq Q_\alpha - \sqrt{s^2 \mathrm{SNR}^2(h)}\right) \tag{1}$$

where $\mathrm{SNR}^2(h) := h^T (J^T F(\mathbf{p}_0)^{-1} J) h$, and $\textit{Type-II}(\phi_n)$ refers to the error of ϕ_n under $\mathbf{p}_n$, and where Q_α is the quantile of order $1-\alpha$ of $\mathcal{N}(0,1)$.

$J := \nabla_\theta f(\theta_0, x_{\text{test}})_{1:(d-1)}$

$F(\mathbf{p}) := \mathrm{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

Theorem 3.1 (Optimal Tests in the LAN Regime). *Let $\alpha \in (0,1)$. If (ϕ_n) is a sequence of tests such that, for every n, ϕ_n is the test with the lowest Type-II error among all tests with Type-I error at most α in testing $\mathbf{p}_0$ vs $\mathbf{p}_n$, then*

$$\textit{Type-II}(\phi_n) \to \mathbb{P}\left(\mathcal{N}(0,1) \leq Q_\alpha - \sqrt{s^2 \mathrm{SNR}^2(h)}\right) \tag{1}$$

where $\mathrm{SNR}^2(h) := h^T (J^T F(\mathbf{p}_0)^{-1} J) h$, *and* $\textit{Type-II}(\phi_n)$ *refers to the error of ϕ_n under $\mathbf{p}_n$, and where Q_α is the quantile of order $1-\alpha$ of $\mathcal{N}(0,1)$.*

Proof: Neyman-Pearson

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

The softmax simplification

Theorem 3.1 (Optimal Tests in the LAN Regime). *Let $\alpha \in (0,1)$. If (ϕ_n) is a sequence of tests such that, for every n, ϕ_n is the test with the lowest Type-II error among all tests with Type-I error at most α in testing $\mathbf{p}_0$ vs $\mathbf{p}_n$, then*

$$\textit{Type-II}(\phi_n) \to \mathbb{P}\left(\mathcal{N}(0,1) \leq Q_\alpha - \sqrt{s^2 \mathrm{SNR}^2(h)}\right) \tag{1}$$

where $\mathrm{SNR}^2(h) := h^T(J^T F(\mathbf{p}_0)^{-1} J)h$, and Type-II$(\phi_n)$ refers to the error of ϕ_n under $\mathbf{p}_n$, and where Q_α is the quantile of order $1-\alpha$ of $\mathcal{N}(0,1)$.

$$F(\mathbf{p}) := \mathrm{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$$

$$J := \nabla_\theta f(\theta_0, x_{\text{test}})_{1:(d-1)}$$

Can we further simplify SNR^2 ?

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

Theorem 3.1 (Optimal Tests in the LAN Regime). *Let $\alpha \in (0,1)$. If (ϕ_n) is a sequence of tests such that, for every n, ϕ_n is the test with the lowest Type-II error among all tests with Type-I error at most α in testing $\mathbf{p}_0$ vs $\mathbf{p}_n$, then*

$$\text{Type-II}(\phi_n) \to \mathbb{P}\left(\mathcal{N}(0,1) \leq Q_\alpha - \sqrt{s^2 \mathrm{SNR}^2(h)}\right) \tag{1}$$

where $\mathrm{SNR}^2(h) := h^T (J^T F(\mathbf{p}_0)^{-1} J) h$, and Type-II$(\phi_n)$ refers to the error of ϕ_n under $\mathbf{p}_n$, and where Q_α is the quantile of order $1-\alpha$ of $\mathcal{N}(0,1)$.

$$F(\mathbf{p}) := \operatorname{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$$

$$\underbrace{r(\theta_{\text{pre}}, w_{<t})}_{\in \mathbb{R}^m} \xrightarrow{\text{linear head } (W,b)} z(\theta, w_{<t}) = W\, r(\theta_{\text{pre}}, w_{<t}) + b \in \mathbb{R}^d, \qquad \theta = (\theta_{\text{pre}}, W, b)$$

$$p_\theta\big(w_t = i \mid w_{<t}\big) = f_i(\theta, w_{<t}) = \frac{\exp\big(z_i(\theta, w_{<t})/\tau\big)}{\sum_{j=1}^{d} \exp\big(z_j(\theta, w_{<t})/\tau\big)}$$

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

Theorem 3.1 (Optimal Tests in the LAN Regime). *Let $\alpha \in (0,1)$. If (ϕ_n) is a sequence of tests such that, for every n, ϕ_n is the test with the lowest Type-II error among all tests with Type-I error at most α in testing $\mathbf{p}_0$ vs $\mathbf{p}_n$, then*

$$\textit{Type-II}(\phi_n) \to \mathbb{P}\left(\mathcal{N}(0,1) \le Q_\alpha - \sqrt{s^2 \mathrm{SNR}^2(h)}\right) \tag{1}$$

where $\mathrm{SNR}^2(h) := h^T (J^T F(\mathbf{p}_0)^{-1} J) h$, and Type-II$(\phi_n)$ refers to the error of ϕ_n under $\mathbf{p}_n$, and where Q_α is the quantile of order $1-\alpha$ of $\mathcal{N}(0,1)$.

$$F(\mathbf{p}) := \mathrm{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$$

$$\underbrace{r(\theta_{\text{pre}}, w_{<t})}_{\in \mathbb{R}^m} \xrightarrow{\text{linear head } (W,b)} z(\theta, w_{<t}) = W\, r(\theta_{\text{pre}}, w_{<t}) + b \in \mathbb{R}^d, \qquad \theta = (\theta_{\text{pre}}, W, b)$$

$$p_\theta(w_t = i \mid w_{<t}) = f_i(\theta, w_{<t}) = \frac{\exp\left(z_i(\theta, w_{<t})/\tau\right)}{\sum_{j=1}^{d} \exp\left(z_j(\theta, w_{<t})/\tau\right)}$$

Lemma 3.2 ($\mathrm{SNR}^2(h)$ in LLMs). *For any $\tau > 0$,*

$$\mathrm{SNR}^2(h) = \frac{1}{\tau^2} h^T \Big(J_z(\theta)^T \underbrace{\Sigma\left(\mathbf{p}^{(\tau)}(\theta)\right)}_{\in \mathbb{R}^{d\times d}} J_z(\theta) \Big) h.$$

$$J_z(\theta) := \nabla_\theta z(\theta, x_{\text{test}}) \in \mathbb{R}^{d\times(q_{\text{pre}}+dm+d)}$$

$$\Sigma\left(\mathbf{p}^{(\tau)}\right) = \mathrm{diag}\left(\mathbf{p}^{(\tau)}\right) - \mathbf{p}^{(\tau)}\left(\mathbf{p}^{(\tau)}\right)^\top \in \mathbb{R}^{d\times d}, \qquad \mathbf{p}^{(\tau)} = \left(\mathbf{p}^{(\tau)}_{1:(d-1)}, 1 - \mathbf{1}^\top \mathbf{p}^{(\tau)}_{1:(d-1)}\right).$$

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

Theorem 3.1 (Optimal Tests in the LAN Regime). *Let $\alpha \in (0,1)$. If (ϕ_n) is a sequence of tests such that, for every n, ϕ_n is the test with the lowest Type-II error among all tests with Type-I error at most α in testing $\mathbf{p}_0$ vs $\mathbf{p}_n$, then*

$$\text{Type-II}(\phi_n) \to \mathbb{P}\left(\mathcal{N}(0,1) \leq Q_\alpha - \sqrt{s^2 \mathrm{SNR}^2(h)}\right) \tag{1}$$

where $\mathrm{SNR}^2(h) := h^T(J^T F(\mathbf{p}_0)^{-1} J)h$, and Type-II$(\phi_n)$ refers to the error of ϕ_n under $\mathbf{p}_n$, and where Q_α is the quantile of order $1-\alpha$ of $\mathcal{N}(0,1)$.

$$F(\mathbf{p}) := \mathrm{diag}(\mathbf{p}) - \mathbf{p}\,\mathbf{p}^\top \in \mathbb{R}^{(d-1)\times(d-1)}$$

$$\underbrace{r(\theta_{\text{pre}}, w_{<t})}_{\in \mathbb{R}^m} \xrightarrow{\text{linear head } (W,b)} z(\theta, w_{<t}) = W\, r(\theta_{\text{pre}}, w_{<t}) + b \in \mathbb{R}^d, \qquad \theta = (\theta_{\text{pre}}, W, b)$$

$$p_\theta(w_t = i \mid w_{<t}) = f_i(\theta, w_{<t}) = \frac{\exp\left(z_i(\theta, w_{<t})/\tau\right)}{\sum_{j=1}^{d} \exp\left(z_j(\theta, w_{<t})/\tau\right)}$$

Lemma 3.2 ($\mathrm{SNR}^2(h)$ in LLMs). *For any $\tau > 0$,*

$$\mathrm{SNR}^2(h) = \frac{1}{\tau^2} h^T \Big(J_z(\theta)^T \underbrace{\Sigma\left(\mathbf{p}^{(\tau)}(\theta)\right)}_{\in \mathbb{R}^{d\times d}} J_z(\theta) \Big) h.$$

$$J_z(\theta) := \nabla_\theta z(\theta, x_{\text{test}}) \in \mathbb{R}^{d\times(q_{\text{pre}}+dm+d)}$$

Proof: Miracle? (Sherman-Morrison + Calculus)

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI

INSTITUT de MATHEMATIQUES de TOULOUSE

Phase transition at low temperature

$$\underbrace{r(\theta_{\text{pre}}, w_{<t})}_{\in \mathbb{R}^m} \xrightarrow{\text{linear head } (W,b)} z(\theta, w_{<t}) = W\, r(\theta_{\text{pre}}, w_{<t}) + b \in \mathbb{R}^d, \qquad \theta = (\theta_{\text{pre}}, W, b) \qquad p_\theta(w_t = i \mid w_{<t}) = f_i(\theta, w_{<t}) = \frac{\exp\big(z_i(\theta, w_{<t})/\tau\big)}{\sum_{j=1}^{d} \exp\big(z_j(\theta, w_{<t})/\tau\big)}$$

$$M := \Big\{ i : z_i(\theta, x) = \max_{1 \le j \le d} z_j(\theta, x) \Big\}, \qquad k := |M|$$

Theorem 3.3 (Phase Transition). *Depending on the value of* k*,*

- *If* $k = 1$*, then* $\mathrm{SNR}^2(h) \to 0$ *as* $\tau \to 0$.
- *If* $k \geq 2$*, and if* $h^T \left(J_z^T \Sigma_M J_z \right) h \neq 0$*, then* $\mathrm{SNR}^2(h) \to +\infty$ *as* $\tau \to 0$.

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI

INSTITUT de MATHEMATIQUES de TOULOUSE

$$\underbrace{r(\theta_{\text{pre}}, w_{<t})}_{\in \mathbb{R}^m} \xrightarrow{\text{linear head } (W,b)} z(\theta, w_{<t}) = W\, r(\theta_{\text{pre}}, w_{<t}) + b \;\in\; \mathbb{R}^d, \qquad \theta = (\theta_{\text{pre}}, W, b) \qquad p_\theta\big(w_t = i \mid w_{<t}\big) \;=\; f_i(\theta, w_{<t}) \;=\; \frac{\exp\big(z_i(\theta, w_{<t})/\tau\big)}{\sum_{j=1}^{d} \exp\big(z_j(\theta, w_{<t})/\tau\big)}$$

$$M := \Big\{ i : z_i(\theta, x) = \max_{1 \le j \le d} z_j(\theta, x) \Big\}, \qquad k := |M|$$

Theorem 3.3 (Phase Transition). *Depending on the value of k,*

- *If $k = 1$, then $\mathrm{SNR}^2(h) \to 0$ as $\tau \to 0$.*
- *If $k \geq 2$, and if $h^T \left(J_z^T \Sigma_{\mathcal{M}} J_z\right) h \neq 0$, then $\mathrm{SNR}^2(h) \to +\infty$ as $\tau \to 0$.*

True for Lebesgue-almost-any h.

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI

INSTITUT de MATHEMATIQUES de TOULOUSE

$$\underbrace{r(\theta_{\text{pre}}, w_{<t})}_{\in \mathbb{R}^m} \xrightarrow{\text{linear head } (W,b)} z(\theta, w_{<t}) = W\, r(\theta_{\text{pre}}, w_{<t}) + b \in \mathbb{R}^d, \qquad \theta = (\theta_{\text{pre}}, W, b) \qquad p_\theta\big(w_t = i \mid w_{<t}\big) = f_i(\theta, w_{<t}) = \frac{\exp\big(z_i(\theta, w_{<t})/\tau\big)}{\sum_{j=1}^{d} \exp\big(z_j(\theta, w_{<t})/\tau\big)}$$

$$M := \Big\{ i : z_i(\theta, x) = \max_{1 \le j \le d} z_j(\theta, x) \Big\}, \qquad k := |M|$$

Theorem 3.3 (Phase Transition). *Depending on the value of k,*

- *If $k = 1$, then $\mathrm{SNR}^2(h) \to 0$ as $\tau \to 0$.*
- *If $k \geq 2$, and if $h^T \left(J_z^T \Sigma_{\mathcal{M}} J_z\right) h \neq 0$, then $\mathrm{SNR}^2(h) \to +\infty$ as $\tau \to 0$.*

True for Lebesgue-almost-any h.

Proof: $tr(J_x^T \Sigma_{\mathcal{M}} J_z) > 0$ thanks to the head contribution.

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

B3IT: the practical algorithm

Theorem 3.3 (Phase Transition). *Depending on the value of k,*

- *If* $k = 1$, *then* $\mathrm{SNR}^2(h) \to 0$ *as* $\tau \to 0$.
- *If* $k \geq 2$, *and if* $h^T \left(J_z^T \Sigma_{\mathcal{M}} J_z\right) h \neq 0$, *then* $\mathrm{SNR}^2(h) \to +\infty$ *as* $\tau \to 0$.

Border Input (BI). A prompt for which at least two tokens are tied for the highest logprob.

$$x \text{ is a Border Input} \iff k(x) = |M(x)| \geq 2.$$

Initialization:

$$\text{For each candidate } x: \text{ sample } m \text{ first tokens at } \tau = 0, \quad Y_1, \ldots, Y_m \sim f(\theta_0, x)\,.$$

$$x \text{ is a BI} \Longleftarrow |\{Y_1, \ldots, Y_m\}| > 1 \quad \text{(more than one distinct output)}\,.$$

Detection:

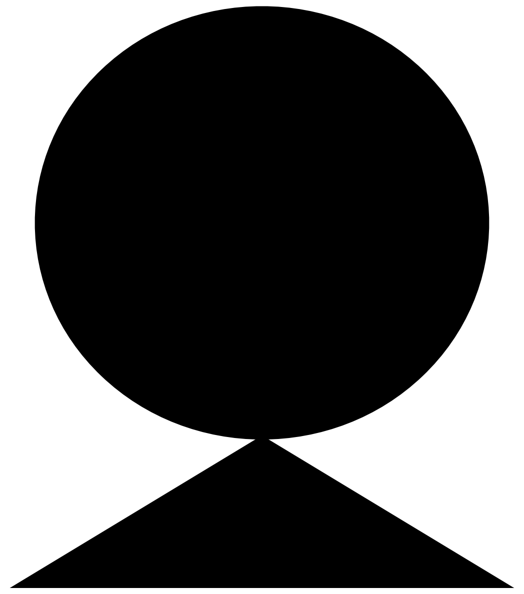
$$\widehat{\mathrm{TV}}\left(\hat{\mathbf{p}}_1(x), \hat{\mathbf{p}}_2(x)\right) > c \implies \text{change detected.}$$

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

In practice, Border Inputs are easy to find

In real life, border inputs should be hard to find.



But we managed to find many rather easily.

78% of endpoints have at least one Border Input.

Found cheaply: $< 1{,}500$ requests for a majority of production APIs.

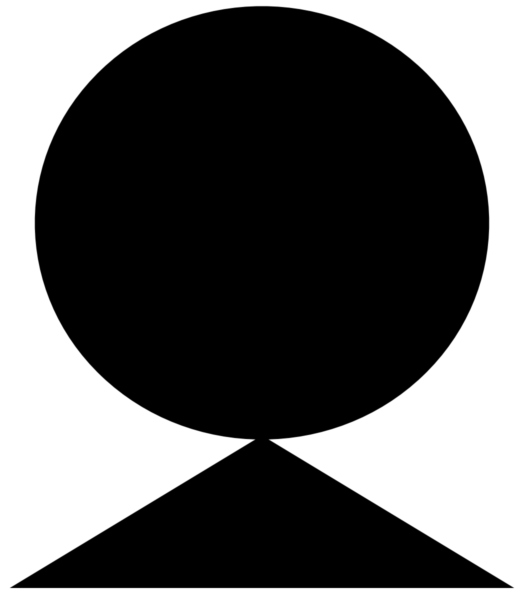
Hypothesis: limited floating-point precision & inference-time non-determinism.

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

In practice, Border Inputs are easy to find

In real life, border inputs should be hard to find.



But we managed to find many rather easily.

78% of endpoints have at least one Border Input.

Found cheaply: $< 1{,}500$ requests for a majority of production APIs.

	1k prompts	2k prompts	2k prompts + reasoning
≥ 1 BIs	95/131 (73%)	99/131 (76%)	102/131 (78%)
≥ 2 BIs	94/131 (72%)	97/131 (74%)	98/131 (75%)
≥ 3 BIs	90/131 (69%)	93/131 (71%)	94/131 (72%)
≥ 4 BIs	87/131 (66%)	92/131 (70%)	93/131 (71%)
≥ 5 BIs	86/131 (66%)	90/131 (69%)	91/131 (69%)

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

```
[BIs: 206, 3000 reqs                    ]  model: z-ai/glm-4.7-flash  provider: phala
[BIs: 189, 1005 reqs                    ]  model: qwen/qwen-2.5-7b-instruct  provider: atlas-cloud/fp8
[BIs: 157, 999 reqs                     ]  model: deepseek/deepseek-v3.2  provider: deepinfra/fp4
[BIs: 136, 2533 reqs                    ]  model: qwen/qwen3-coder-30b-a3b-instruct  provider: siliconflow/fp8
[BIs: 121, 1056 reqs                    ]  model: meta-llama/llama-4-scout  provider: groq
[BIs: 119, 447 reqs                     ]  model: qwen/qwen3.5-9b  provider: venice/fp8
[BIs: 107, 1445 reqs                    ]  model: meituan/longcat-flash-chat  provider: atlas-cloud/fp8
[BIs: 106, 966 reqs                     ]  model: deepseek/deepseek-chat-v3-0324  provider: deepinfra/fp4
[BIs: 95, 2928 reqs                     ]  model: qwen/qwen3-235b-a22b-2507  provider: deepinfra/fp8
[BIs: 86, 978 reqs                      ]  model: qwen/qwen-turbo  provider: alibaba
[BIs: 72, 942 reqs                      ]  model: deepseek/deepseek-v3.2  provider: atlas-cloud/fp8
[BIs: 71, 1767 reqs                     ]  model: qwen/qwen3-30b-a3b-instruct-2507  provider: siliconflow/fp8
[BIs: 61, 945 reqs                      ]  model: meta-llama/llama-4-scout  provider: deepinfra/fp8
[BIs: 60, 966 reqs                      ]  model: mistralai/voxtral-small-24b-2507  provider: mistral
[BIs: 58, 900 reqs                      ]  model: deepseek/deepseek-v3.2-exp  provider: novita/fp8
[BIs: 54, 492 reqs                      ]  model: deepseek/deepseek-v3.2  provider: chutes/fp8
[BIs: 54, 900 reqs                      ]  model: deepseek/deepseek-chat-v3.1  provider: deepinfra/fp4
[BIs: 52, 1014 reqs, reasoning: disabled ]  model: z-ai/glm-4.7-flash  provider: deepinfra/bf16
[BIs: 51, 1530 reqs                     ]  model: z-ai/glm-4.7-flash  provider: z-ai
[BIs: 51, 3001 reqs                     ]  model: qwen/qwen3-8b  provider: atlas-cloud/fp8
[BIs: 49, 2999 reqs                     ]  model: qwen/qwen3-8b  provider: alibaba
[BIs: 45, 807 reqs                      ]  model: deepseek/deepseek-v3.1-terminus  provider: deepinfra/fp4
[BIs: 45, 939 reqs                      ]  model: cohere/command-r-08-2024  provider: cohere
[BIs: 43, 3000 reqs                     ]  model: meta-llama/llama-3.1-70b-instruct  provider: deepinfra/turbo
[BIs: 41, 297 reqs                      ]  model: tencent/hunyuan-a13b-instruct  provider: siliconflow/fp8
[BIs: 40, 468 reqs                      ]  model: deepseek/deepseek-v3.2-exp  provider: atlas-cloud/fp8
[BIs: 40, 954 reqs                      ]  model: mistralai/ministral-8b-2512  provider: mistral
[BIs: 40, 1113 reqs                     ]  model: google/gemma-3-27b-it  provider: parasail/fp8
[BIs: 34, 936 reqs                      ]  model: mistralai/ministral-14b-2512  provider: mistral
[BIs: 33, 2997 reqs                     ]  model: openai/gpt-4o-mini  provider: openai
[BIs: 32, 3000 reqs                     ]  model: openai/gpt-4o-mini-2024-07-18  provider: openai
[BIs: 31, 762 reqs                      ]  model: qwen/qwen3-next-80b-a3b-instruct  provider: alibaba
[BIs: 30, 789 reqs                      ]  model: qwen/qwen3-vl-30b-a3b-instruct  provider: alibaba
[BIs: 29, 2988 reqs                     ]  model: amazon/nova-lite-v1  provider: amazon-bedrock
[BIs: 29, 3000 reqs                     ]  model: google/gemini-2.0-flash-lite-001  provider: google-ai-studio
[BIs: 29, 3000 reqs                     ]  model: sao10k/l3-lunaris-8b  provider: novita/bf16
[BIs: 28, 747 reqs                      ]  model: qwen/qwen3-vl-8b-instruct  provider: novita/fp8
[BIs: 28, 3000 reqs                     ]  model: google/gemini-2.0-flash-lite-001  provider: google-vertex
[BIs: 27, 3000 reqs                     ]  model: mistralai/ministral-8b-2512  provider: nextbit/fp8
[BIs: 26, 3000 reqs                     ]  model: openai/gpt-4o-mini  provider: azure
[BIs: 25, 327 reqs                      ]  model: deepseek/deepseek-v3.2  provider: akashml/fp8
[BIs: 25, 597 reqs                      ]  model: meta-llama/llama-3.3-70b-instruct  provider: novita/bf16
[BIs: 25, 981 reqs                      ]  model: meta-llama/llama-3-8b-instruct  provider: novita/bf16
[BIs: 24, 522 reqs                      ]  model: qwen/qwq-32b  provider: siliconflow/fp8
[BIs: 24, 618 reqs, reasoning: disabled  ]  model: qwen/qwen3-30b-a3b  provider: novita/fp8
[BIs: 23, 708 reqs                      ]  model: meta-llama/llama-3.1-8b-instruct  provider: nebius/fp8
[BIs: 21, 807 reqs                      ]  model: qwen/qwen3-coder-30b-a3b-instruct  provider: novita/fp8
[BIs: 20, 3000 reqs                     ]  model: meta-llama/llama-3-8b-instruct  provider: deepinfra/bf16
[BIs: 20, 3000 reqs                     ]  model: qwen/qwen-2.5-7b-instruct  provider: together/fp8
[BIs: 19, 616 reqs                      ]  model: z-ai/glm-4.5-air  provider: siliconflow/fp8
[BIs: 19, 741 reqs                      ]  model: mistralai/ministral-3b-2512  provider: mistral
[BIs: 19, 3000 reqs                     ]  model: google/gemini-2.5-flash-lite  provider: google-vertex
[BIs: 18, 861 reqs                      ]  model: qwen/qwen3-vl-30b-a3b-instruct  provider: deepinfra/fp8
[BIs: 18, 2457 reqs                     ]  model: google/gemma-3-27b-it  provider: deepinfra/fp8
[BIs: 18, 3000 reqs                     ]  model: mistralai/ministral-3b-2512  provider: nextbit/fp8
[BIs: 17, 294 reqs                      ]  model: z-ai/glm-4.7-flash  provider: novita/bf16
[BIs: 17, 762 reqs                      ]  model: mistralai/mistral-small-3.2-24b-instruct  provider: deepinfra
[BIs: 17, 5997 reqs                     ]  model: qwen/qwen3.5-flash-02-23  provider: alibaba
[BIs: 16, 3000 reqs                     ]  model: mistralai/ministral-14b-2512  provider: nextbit/fp8
[BIs: 15, 2970 reqs                     ]  model: thedrummer/unslopnemo-12b  provider: nextbit/fp8
[BIs: 13, 3000 reqs                     ]  model: meta-llama/llama-3.1-70b-instruct  provider: deepinfra/base
[BIs: 13, 3000 reqs                     ]  model: qwen/qwen3.5-9b  provider: together
[BIs: 12, 459 reqs                      ]  model: qwen/qwen3-30b-a3b-instruct-2507  provider: atlas-cloud/fp8
[BIs: 12, 3000 reqs, reasoning: budget=4 ]  model: alibaba/tongyi-deepresearch-30b-a3b  provider: atlas-cloud/fp
[BIs: 12, 3000 reqs                     ]  model: meta-llama/llama-3.2-11b-vision-instruct  provider: deepinfra
[BIs: 12, 3000 reqs                     ]  model: mistralai/mistral-small-3.2-24b-instruct  provider: parasail/
[BIs: 11, 384 reqs                      ]  model: mistralai/mistral-small-3.2-24b-instruct  provider: venice/fp
[BIs: 11, 852 reqs                      ]  model: meta-llama/llama-3.1-8b-instruct  provider: novita/fp8
[BIs: 11, 3000 reqs                     ]  model: mistralai/mistral-small-3.1-24b-instruct  provider: cloudflar
[BIs: 10, 3000 reqs                     ]  model: meta-llama/llama-3-8b-instruct  provider: together/int4
[BIs: 10, 3000 reqs                     ]  model: mistralai/mistral-nemo  provider: novita/fp8
[BIs: 10, 3000 reqs                     ]  model: stepfun/step-3.5-flash  provider: stepfun/fp8
[BIs: 9, 375 reqs                       ]  model: qwen/qwen3-30b-a3b-thinking-2507  provider: siliconflow/fp8
[BIs: 9, 600 reqs                       ]  model: stepfun/step-3.5-flash  provider: deepinfra/fp8
[BIs: 9, 2016 reqs                      ]  model: google/gemma-3-27b-it  provider: novita/bf16
[BIs: 9, 3000 reqs                      ]  model: qwen/qwen3-vl-32b-instruct  provider: alibaba
[BIs: 8, 600 reqs                       ]  model: qwen/qwen-vl-plus  provider: alibaba
[BIs: 8, 600 reqs                       ]  model: qwen/qwen3-coder-next  provider: chutes/bf16
[BIs: 8, 3000 reqs                      ]  model: google/gemma-3-12b-it  provider: cloudflare
[BIs: 8, 3000 reqs                      ]  model: google/gemma-3-4b-it  provider: deepinfra/bf16
[BIs: 8, 3000 reqs                      ]  model: meta-llama/llama-3.1-8b-instruct  provider: deepinfra/bf16
[BIs: 8, 5997 reqs                      ]  model: amazon/nova-micro-v1  provider: amazon-bedrock
[BIs: 7, 600 reqs                       ]  model: google/gemini-2.0-flash-001  provider: google-vertex
[BIs: 7, 1554 reqs, reasoning: disabled  ]  model: z-ai/glm-4.5-air  provider: novita/bf16
```

```
[BIs: 6, 600 reqs                       ]  model: google/gemini-2.0-flash-001  provider: google-ai-studio
[BIs: 6, 600 reqs                       ]  model: meta-llama/llama-4-maverick  provider: deepinfra/base
[BIs: 6, 3000 reqs                      ]  model: google/gemma-3-12b-it  provider: deepinfra/bf16
[BIs: 6, 6034 reqs                      ]  model: qwen/qwen3-coder-next  provider: ionstream/fp8
[BIs: 5, 3000 reqs                      ]  model: qwen/qwen3-235b-a22b-2507  provider: wandb/bf16
[BIs: 5, 3000 reqs                      ]  model: sao10k/l3-lunaris-8b  provider: deepinfra/turbo
[BIs: 5, 6000 reqs                      ]  model: meta-llama/llama-3.2-1b-instruct  provider: cloudflare
[BIs: 4, 6000 reqs                      ]  model: liquid/lfm-2-24b-a2b  provider: together
```

```
[BIs: 4, 6000 reqs                      ]  model: x-ai/grok-3-mini  provider: xai
[BIs: 3, 6018 reqs                      ]  model: x-ai/grok-3-mini-beta  provider: xai
[BIs: 2, 6000 reqs                      ]  model: meta-llama/llama-3.1-8b-instruct  provider: friendli
[BIs: 2, 6000 reqs                      ]  model: meta-llama/llama-3.2-3b-instruct  provider: cloudflare
[BIs: 2, 6000 reqs                      ]  model: mistralai/mistral-nemo  provider: deepinfra/fp8
[BIs: 2, 6000 reqs                      ]  model: reka/reka-edge  provider: reka/bf16
[BIs: 1, 3477 reqs, reasoning: budget=256]  model: qwen/qwen3-32b  provider: chutes/bf16
[BIs: 1, 6000 reqs                      ]  model: microsoft/phi-4  provider: nextbit/int4
[BIs: 1, 6000 reqs, reasoning: budget=256]  model: qwen/qwen3-32b  provider: groq
[BIs: 1, 6042 reqs                      ]  model: qwen/qwen3-next-80b-a3b-thinking  provider: alibaba
[BIs: 0, 5944 reqs                      ]  model: deepseek/deepseek-chat-v3.1  provider: sambanova/high-throughput
[BIs: 0, 6000 reqs                      ]  model: arcee-ai/trinity-mini  provider: clarifai/bf16
[BIs: 0, 6000 reqs                      ]  model: deepseek/deepseek-v3.2  provider: deepseek
[BIs: 0, 6000 reqs                      ]  model: google/gemini-2.5-flash-lite  provider: google-ai-studio
[BIs: 0, 6000 reqs                      ]  model: google/gemini-2.5-flash-lite-preview-09-2025  provider: google-ai-studio
[BIs: 0, 6000 reqs                      ]  model: google/gemini-2.5-flash-lite-preview-09-2025  provider: google-vertex
[BIs: 0, 6000 reqs                      ]  model: google/gemma-2-9b-it  provider: nebius/fast
[BIs: 0, 6000 reqs                      ]  model: liquid/lfm2-8b-a1b  provider: liquid
[BIs: 0, 6000 reqs                      ]  model: meta-llama/llama-3.1-8b-instruct  provider: cerebras/fp16
[BIs: 0, 6000 reqs                      ]  model: meta-llama/llama-3.1-8b-instruct  provider: sambanova/bf16
[BIs: 0, 6000 reqs                      ]  model: meta-llama/llama-3.3-70b-instruct  provider: nebius/fp8
[BIs: 0, 6000 reqs                      ]  model: microsoft/phi-4  provider: deepinfra/bf16
[BIs: 0, 6000 reqs                      ]  model: mistralai/devstral-small  provider: mistral
[BIs: 0, 6000 reqs                      ]  model: mistralai/mistral-7b-instruct-v0.1  provider: cloudflare
[BIs: 0, 6000 reqs                      ]  model: nousresearch/hermes-2-pro-llama-3-8b  provider: novita/fp16
[BIs: 0, 6000 reqs                      ]  model: nousresearch/hermes-3-llama-3.1-70b  provider: deepinfra/fp8
[BIs: 0, 6000 reqs                      ]  model: nvidia/nemotron-nano-12b-v2-vl  provider: deepinfra/fp8
[BIs: 0, 6000 reqs                      ]  model: nvidia/nemotron-nano-9b-v2  provider: deepinfra/bf16
[BIs: 0, 6000 reqs                      ]  model: qwen/qwen-2.5-72b-instruct  provider: novita/bf16
[BIs: 0, 6000 reqs, reasoning: budget=1  ]  model: qwen/qwen3-14b  provider: deepinfra/fp8
[BIs: 0, 6000 reqs, reasoning: budget=128]  model: qwen/qwen3-14b  provider: nextbit/int4
[BIs: 0, 6000 reqs, reasoning: budget=256]  model: qwen/qwen3-30b-a3b  provider: deepinfra/fp8
[BIs: 0, 6000 reqs, reasoning: budget=512]  model: qwen/qwen3-30b-a3b  provider: friendli
[BIs: 0, 6000 reqs                      ]  model: qwen/qwen3-30b-a3b-instruct-2507  provider: wandb/bf16
[BIs: 0, 6000 reqs                      ]  model: qwen/qwen3-30b-a3b-thinking-2507  provider: nebius/fp8
[BIs: 0, 6000 reqs, reasoning: budget=256]  model: qwen/qwen3-32b  provider: deepinfra/fp8
[BIs: 0, 6000 reqs                      ]  model: qwen/qwen3-vl-8b-instruct  provider: alibaba
[BIs: 0 (hidden reasoning)              ]  model: openai/gpt-oss-20b  provider: amazon-bedrock
[BIs: 0 (hidden reasoning)              ]  model: openai/gpt-oss-120b  provider: amazon-bedrock
```

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

In-vitro experiments

TinyChange Benchmark

Qwen/Qwen2.5-0.5B-Instruct

Qwen/Qwen2.5-7B-Instruct

google/gemma-3-1b-it

google/gemma-2-9b-it

microsoft/Phi-4-mini-instruct

mistralai/Mistral-7B-Instruct-v0.3

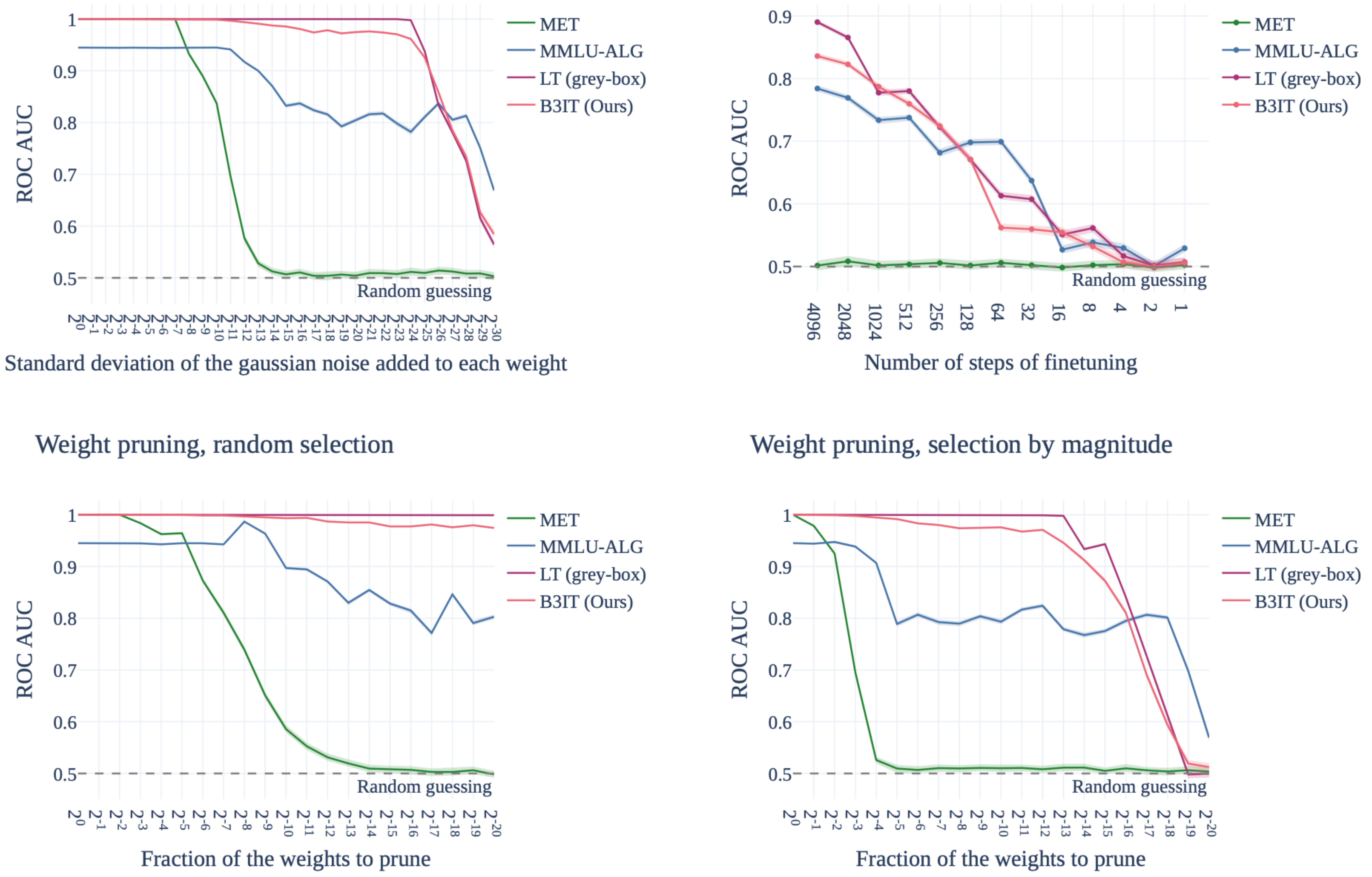
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

meta-llama/Llama-3.1-8B-Instruct

allenai/OLMo-2-1124-7B-Instruct

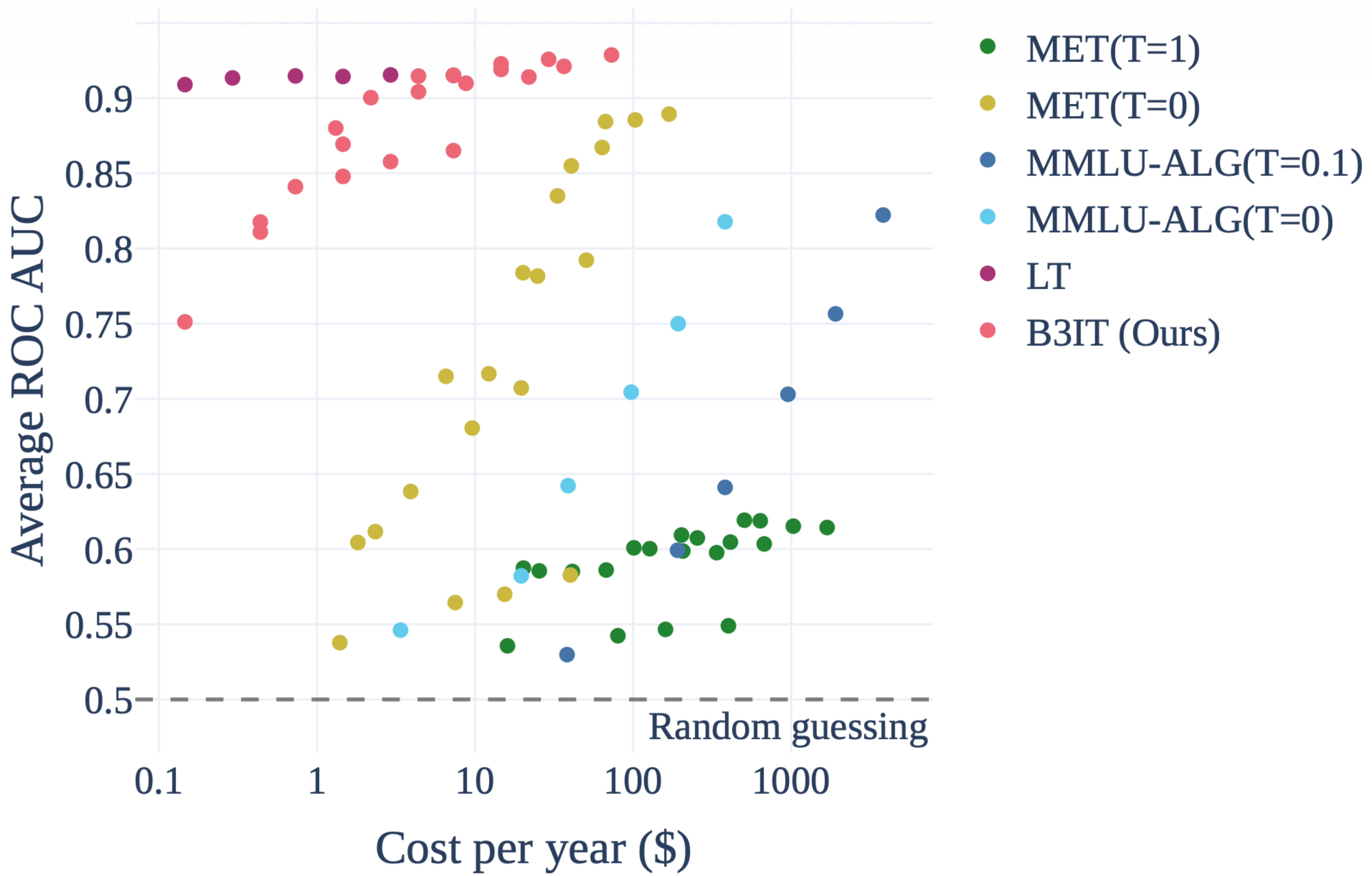
Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

In-vitro experiments



Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

In-vitro experiments



Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI

INSTITUT de MATHEMATIQUES de TOULOUSE

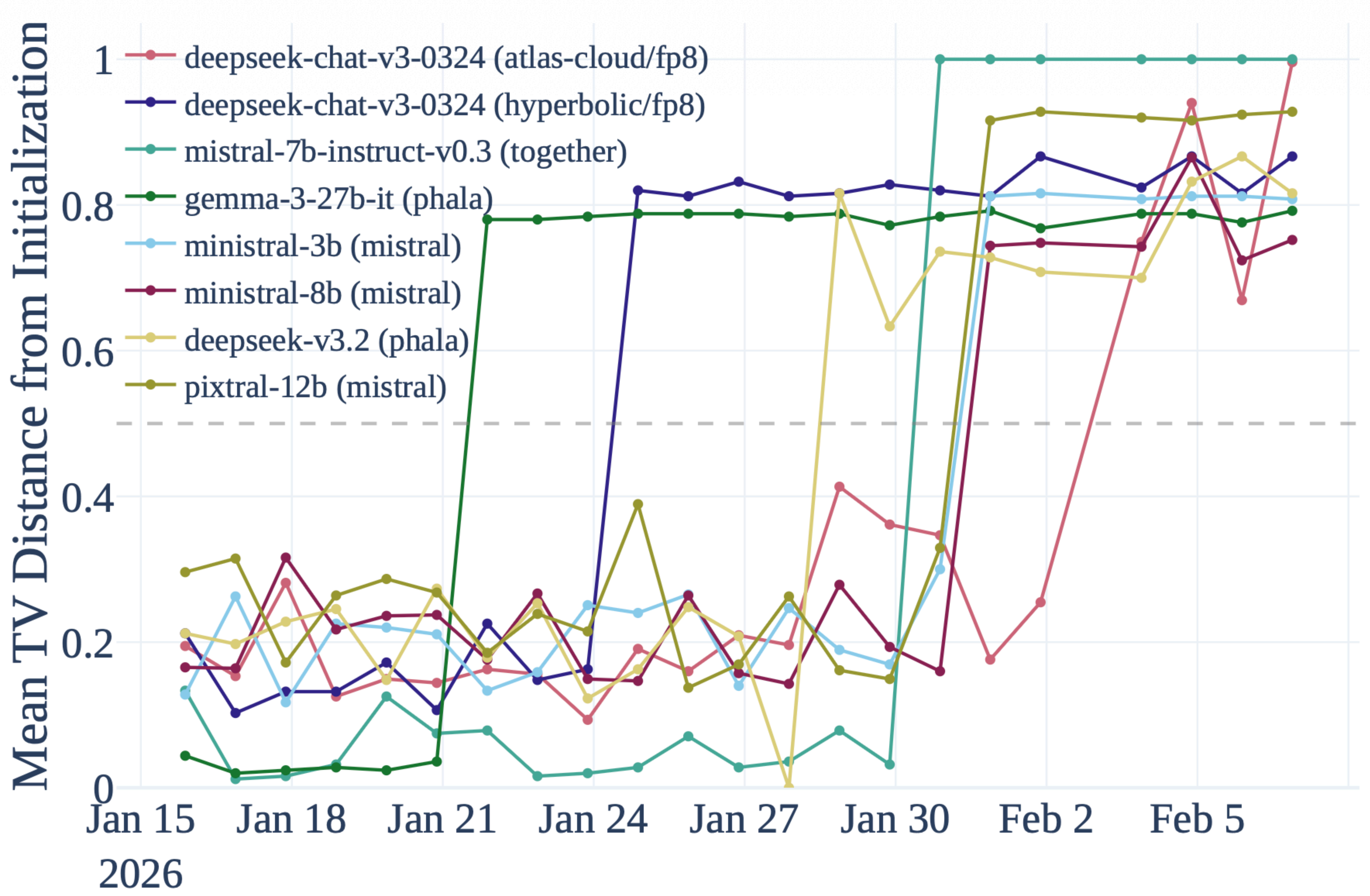
In-vivo detections

```
[BIs: 61, 1720 reqs ]  model: mistralai/mistral-nemo  provider: mistral
[BIs: 60, 540 reqs  ]  model: mistralai/mistral-nemo  provider: azure
[BIs: 60, 780 reqs  ]  model: deepseek/deepseek-v3.2  provider: parasail/fp8
[BIs: 60, 880 reqs  ]  model: deepseek/deepseek-chat-v3-0324  provider: fireworks
[BIs: 60, 960 reqs  ]  model: relace/relace-search  provider: relace/bf16
[BIs: 60, 1120 reqs ]  model: mistralai/ministral-3b  provider: mistral
[BIs: 60, 2440 reqs ]  model: qwen/qwen3-vl-30b-a3b-instruct  provider: fireworks
[BIs: 60, 2620 reqs ]  model: google/gemma-3-12b-it  provider: chutes/bf16
[BIs: 60, 4980 reqs ]  model: qwen/qwen3-235b-a22b-2507  provider: wandb/bf16
[BIs: 60, 7260 reqs ]  model: mistralai/mistral-nemo  provider: chutes/bf16
[BIs: 60, 10320 reqs]  model: google/gemma-3-4b-it  provider: deepinfra/bf16
[BIs: 60, 10780 reqs]  model: deepseek/deepseek-chat-v3.1  provider: sambanova/fp8
[BIs: 60, 12580 reqs]  model: qwen/qwen3-235b-a22b-2507  provider: crusoe/bf16
[BIs: 60, 14860 reqs]  model: qwen/qwen-2.5-coder-32b-instruct  provider: chutes/fp8
[BIs: 59, 1320 reqs ]  model: openai/gpt-4o  provider: openai
[BIs: 59, 1340 reqs ]  model: deepseek/deepseek-chat-v3-0324  provider: hyperbolic/fp8
[BIs: 59, 2340 reqs ]  model: kwaipilot/kat-coder-pro  provider: streamlake/fp16
[BIs: 59, 5880 reqs ]  model: openai/gpt-4o-mini  provider: openai
[BIs: 59, 6740 reqs ]  model: mistralai/devstral-small-2505  provider: deepinfra/bf16
[BIs: 59, 13400 reqs]  model: mistralai/mistral-small-3.2-24b-instruct  provider: parasail/bf16
[BIs: 58, 580 reqs  ]  model: deepseek/deepseek-chat-v3-0324  provider: atlas-cloud/fp8
[BIs: 58, 1120 reqs ]  model: cohere/command-r-08-2024  provider: cohere
[BIs: 58, 1340 reqs ]  model: mistralai/ministral-8b-2512  provider: mistral
[BIs: 58, 1360 reqs ]  model: mistralai/pixtral-12b  provider: mistral
[BIs: 58, 4520 reqs ]  model: moonshotai/kimi-k2-0905:exacto  provider: moonshotai
[BIs: 58, 6920 reqs ]  model: google/gemma-3-27b-it  provider: chutes/bf16
[BIs: 58, 7480 reqs ]  model: openai/gpt-4o-mini  provider: azure
[BIs: 58, 11680 reqs]  model: amazon/nova-micro-v1  provider: amazon-bedrock
[BIs: 57, 1200 reqs ]  model: mistralai/ministral-3b-2512  provider: mistral
[BIs: 57, 1400 reqs ]  model: deepseek/deepseek-v3.2  provider: phala
[BIs: 57, 2040 reqs ]  model: google/gemma-3-12b-it  provider: crusoe/bf16
[BIs: 56, 1500 reqs ]  model: mistralai/ministral-8b  provider: mistral
[BIs: 56, 10480 reqs]  model: amazon/nova-lite-v1  provider: amazon-bedrock
[BIs: 56, 15000 reqs]  model: microsoft/phi-4  provider: deepinfra/bf16
[BIs: 52, 3620 reqs ]  model: inflection/inflection-3-pi  provider: inflection
[BIs: 50, 4720 reqs ]  model: anthropic/claude-sonnet-4.5  provider: anthropic
[BIs: 43, 1740 reqs ]  model: meta-llama/llama-guard-2-8b  provider: together
[BIs: 42, 6040 reqs ]  model: google/gemma-3-12b-it  provider: novita/bf16
[BIs: 40, 15000 reqs]  model: google/gemma-3-12b-it  provider: deepinfra/bf16
[BIs: 36, 15000 reqs]  model: anthropic/claude-haiku-4.5  provider: anthropic
[BIs: 36, 15000 reqs]  model: microsoft/phi-4-multimodal-instruct  provider: deepinfra/bf16
[BIs: 35, 8840 reqs ]  model: mistralai/mistral-nemo  provider: deepinfra/fp8
```

```
[BIs: 28, 4920 reqs ]  model: qwen/qwen-2.5-72b-instruct  provider: hyperbolic/bf16
[BIs: 26, 15000 reqs]  model: mistralai/mistral-7b-instruct-v0.3  provider: together
[BIs: 25, 15000 reqs]  model: nousresearch/hermes-2-pro-llama-3-8b  provider: nextbit/int4
[BIs: 23, 600 reqs  ]  model: deepseek/deepseek-chat-v3-0324  provider: baseten/fp8
[BIs: 21, 15000 reqs]  model: qwen/qwen3-235b-a22b-2507  provider: cerebras
[BIs: 19, 15000 reqs]  model: qwen/qwen3-vl-235b-a22b-instruct  provider: fireworks
[BIs: 14, 6060 reqs ]  model: alibaba/tongyi-deepresearch-30b-a3b  provider: ncompass/bf16
[BIs: 11, 2800 reqs ]  model: google/gemma-3-27b-it  provider: phala
[BIs: 10, 15000 reqs]  model: google/gemma-3-4b-it  provider: chutes
[BIs: 7, 15000 reqs ]  model: mistralai/devstral-medium  provider: mistral
[BIs: 1, 15000 reqs ]  model: microsoft/phi-4  provider: nextbit/int4
[BIs: 0, 0 reqs     ]  model: thudm/glm-4.1v-9b-thinking  provider: novita/bf16
[BIs: 0, 120 reqs   ]  model: qwen/qwen3-coder  provider: baseten/fp8
[BIs: 0, 8920 reqs  ]  model: qwen/qwen3-32b  provider: friendli
[BIs: 0, 10240 reqs ]  model: qwen/qwen3-30b-a3b  provider: friendli
[BIs: 0, 14980 reqs ]  model: anthropic/claude-sonnet-4.5  provider: amazon-bedrock
[BIs: 0, 14980 reqs ]  model: baidu/ernie-4.5-21b-a3b  provider: novita/bf16
[BIs: 0, 14980 reqs ]  model: cohere/command-r7b-12-2024  provider: cohere
[BIs: 0, 14980 reqs ]  model: z-ai/glm-4.6:exacto  provider: z-ai
[BIs: 0, 15000 reqs ]  model: deepseek/deepseek-r1-distill-llama-70b  provider: chutes/bf16
[BIs: 0, 15000 reqs ]  model: deepseek/deepseek-v3.2  provider: atlas-cloud/fp8
[BIs: 0, 15000 reqs ]  model: deepseek/deepseek-v3.2  provider: deepseek
[BIs: 0, 15000 reqs ]  model: google/gemini-2.5-flash-lite-preview-09-2025  provider: google-ai-studio
[BIs: 0, 15000 reqs ]  model: google/gemini-2.5-flash-lite-preview-09-2025  provider: google-vertex
[BIs: 0, 15000 reqs ]  model: google/gemini-3-flash-preview  provider: google-ai-studio
[BIs: 0, 15000 reqs ]  model: google/gemini-3-flash-preview  provider: google-vertex
[BIs: 0, 15000 reqs ]  model: google/gemma-2-9b-it  provider: nebius/fast
[BIs: 0, 15000 reqs ]  model: google/gemma-3-27b-it  provider: ncompass/fp8
[BIs: 0, 15000 reqs ]  model: liquid/lfm-2.2-6b  provider: liquid
[BIs: 0, 15000 reqs ]  model: liquid/lfm2-8b-a1b  provider: liquid
[BIs: 0, 15000 reqs ]  model: meta-llama/llama-3.1-405b  provider: hyperbolic/bf16
[BIs: 0, 15000 reqs ]  model: mistralai/mistral-7b-instruct-v0.1  provider: cloudflare
[BIs: 0, 15000 reqs ]  model: moonshotai/kimi-k2-thinking  provider: moonshotai/int4
[BIs: 0, 15000 reqs ]  model: nousresearch/hermes-2-pro-llama-3-8b  provider: novita/fp16
[BIs: 0, 15000 reqs ]  model: openai/gpt-oss-120b  provider: amazon-bedrock
[BIs: 0, 15000 reqs ]  model: openai/gpt-oss-20b  provider: amazon-bedrock
[BIs: 0, 15000 reqs ]  model: prime-intellect/intellect-3  provider: nebius/fp8
[BIs: 0, 15000 reqs ]  model: qwen/qwen3-14b  provider: chutes/bf16
[BIs: 0, 15000 reqs ]  model: qwen/qwen3-14b  provider: nextbit/int4
[BIs: 0, 15000 reqs ]  model: qwen/qwen3-14b  provider: deepinfra/fp8
[BIs: 0, 15000 reqs ]  model: qwen/qwen3-30b-a3b-thinking-2507  provider: cloudflare
[BIs: 0, 15000 reqs ]  model: qwen/qwen3-32b  provider: nebius/base
[BIs: 0, 15000 reqs ]  model: qwen/qwen3-32b  provider: groq
[BIs: 0, 15000 reqs ]  model: qwen/qwen3-32b  provider: cerebras
[BIs: 0, 15000 reqs ]  model: qwen/qwen3-32b  provider: sambanova
[BIs: 0, 15000 reqs ]  model: qwen/qwen3-8b  provider: fireworks
[BIs: 0, 15000 reqs ]  model: z-ai/glm-4-32b  provider: z-ai
[BIs: 0, 15000 reqs ]  model: z-ai/glm-4.5  provider: wandb/bf16
[BIs: 0, 15000 reqs ]  model: z-ai/glm-4.5-air  provider: chutes/bf16
[BIs: 0, 15000 reqs ]  model: z-ai/glm-4.6  provider: mancer/fp8
[BIs: 0, 15000 reqs ]  model: z-ai/glm-4.7  provider: mancer/fp8
```

Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

In-vivo detections



Source: Chauvin et al., Token-Efficient Change Detection in LLM APIs, ICML 2026

ANITI INSTITUT de MATHEMATIQUES de TOULOUSE

Future work directions

- Can we leverage Border Inputs to define a “distance” between models ?
- Can we “robustify” the detection to benign changes ?
- Can we “prove” a malicious change ?
- Better search strategy for Border Inputs.

Thank You!

ANITI

INSTITUT de MATHEMATIQUES de TOULOUSE