A Geometric Analysis of Small-sized Language Model Hallucinations

APORIA | ICML 2026

Elia Onofri
elia.onofri@kaust.edu.sa

CRI-Lab
King Abdullah University of Science and Technology (KAUST)

ICML 2026 · Seoul

The Team

Emanuele Ricco
CRI-Lab, KAUST

Elia Onofri
CRI-Lab, KAUST

Roberto Di Pietro
CRI-Lab, KAUST

Lorenzo Cima
IIT–CNR, Italy

Stefano Cresci
IIT–CNR, Italy

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

ISTITUTO
DI INFORMATICA
E TELEMATICA
CNR

Fluent, confident . . . and wrong

“Who won the Wimbledon men’s singles title in 2022?”

→

“**Carlos Alcaraz** won the Wimbledon men’s singles title in 2022. He defeated Novak Djokovic in the final [. . .]”

fluent ✓ **confident ✓** **correct ✗**

Hallucination

A fluent, confident, but false answer.

Why it matters

A core obstacle to deploying LLMs where reliability is at a premium.

The same question, asked many times

sometimes right
sometimes wrong

Key observation

The model is right some of the time $\Rightarrow$ the knowledge is there.
The failure is *not* missing knowledge — it is something else.

[*The effect is present across all LLMs; most apparent in small-sized models (a few billion parameters).*]

Our claim

Hallucination is *retrieval instability*, not missing knowledge.

✗ a gap in what the model *knows* vs ✓ instability in *retrieving* what it knows

[. . . *an internal failure whose inner workings remain difficult to explain — so we need a way to see it.*]

APORIA: seeing the instability

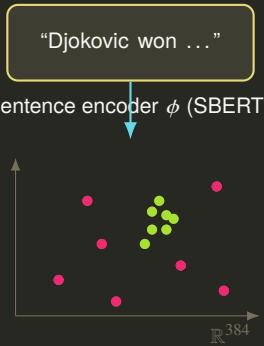
The name

APORIA — *Aggregate Prompt-wise Observation Retrieving Instability via Asymmetry*.

From Socratic philosophy

Aporía — the puzzlement-in-contradiction when a confident claim, probed again and again, contradicts itself.

Exactly what a hallucination is.

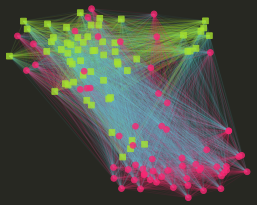


similar meaning → nearby points
comparing answers = measuring distance

The pattern: cluster vs. scatter

Raw embedding space

[The image is a t-SNE approximation of real data]



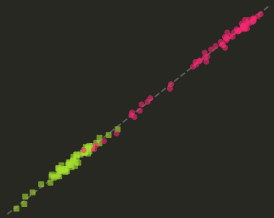
faint, overlapping

Fisher

$\Longrightarrow$

projection

Fisher projection (1D)



What we measure

Distance distributions within and between classes, compared with the Wasserstein distance.

On SOCRATES-300K

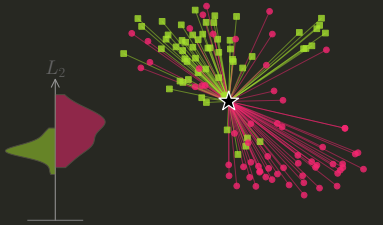
separation ratio

≈ 1 (raw) $\longrightarrow$ > 7 (Fisher)

APORIA-LP: propagation via geometry

Raw embedding space

[The image is a t-SNE approximation of real data]

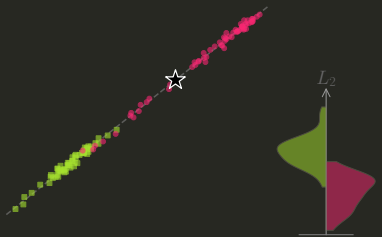


Fisher

$\Longrightarrow$

projection

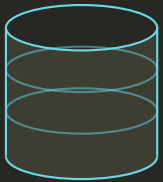
Fisher projection (1D)



Result

From only a small labelled set, APORIA-LP automatically tags genuine vs. hallucinated, reaching F1 $> 90\%$ across ten models.

Takeaway & SOCRATES-300K release



SOCRATES-300K

- **300 000** labelled responses
- 200 prompts × 150 generations × 10 models
- dataset + generation + reproduction code

data: doi.org/10.5281/zenodo.20256462

code: github.com/eOnofri04/APORIA

Hallucination becomes a measurable, predictable property —
a step toward LLMs we can ultimately trust.

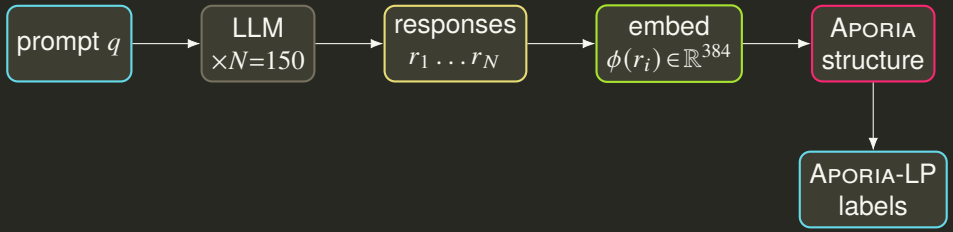
Thank you!

elia.onofri@kaust.edu.sa

[*Feel free to explore the other slides for other insights on our work.*]

[*Slides available on* www.eliaonofri.it]

A.1 | Method pipeline



[*LLM-as-a-judge (Claude) provides genuine/hallucinated labels; the encoder ϕ is fixed and model-independent.*]

A.2 | The SOCRATES-300K dataset

- Time-sensitive factual prompts
- 10 small-sized LLMs (7–32B), 4-bit, identical decoding
- 150 generations per (model, prompt)
- LLM-as-a-judge labels: *Genuine* / *Hallucinated* / *Unknown*
- 231 473 responses analysed after filtering

Model	Hallucination Rate [%]			Dropped Prompts [%]		
	2020	2022	Global	2020	2022	Global
mistral-7B	**86.1** (7.8)	**89.0** (6.2)	**87.6** (7.1)	28.0	26.0	27.0
deepseek-7B	84.5 (11.6)	85.0 (9.7)	84.8 (10.7)	29.0	35.0	32.0
llama3-8B	80.4 (13.0)	82.5 (11.6)	81.4 (12.3)	21.0	16.0	18.5
gemma2-9B	85.6 (6.8)	86.2 (6.6)	85.9 (6.7)	24.0	5.0	14.5
yi1.5-9B	85.9 (11.1)	88.1 (8.5)	86.9 (10.0)	36.0	47.0	41.5
solar1-11B	72.3 (13.8)	74.5 (14.6)	73.4 (14.2)	17.0	7.0	12.0
phi4-14B	64.7 (25.1)	68.4 (21.5)	66.4 (23.5)	13.0	28.0	20.5
qwen2.5-14B	77.6 (16.3)	78.5 (14.7)	78.0 (15.6)	11.0	29.0	20.0
gemma2-27B	83.5 (11.7)	88.6 (5.8)	86.3 (9.3)	24.0	6.0	15.0
qwen2.5-32B	75.0 (17.0)	82.5 (12.4)	78.5 (15.5)	16.0	28.0	22.0
Average	79.6 (13.4)	82.3 (11.2)	80.9 (12.5)	21.9	22.7	22.3

A.3 | Distances, Wasserstein, Fisher

Three empirical distance distributions per prompt:

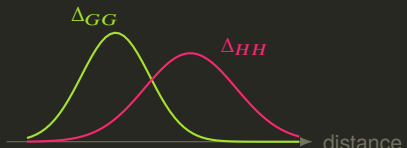
$$\Delta_{GG}, \quad \Delta_{HH}, \quad \Delta_{GH}.$$

Separation scored by the 1-D Wasserstein distance vs. a label-permutation null H_0.

Fisher direction

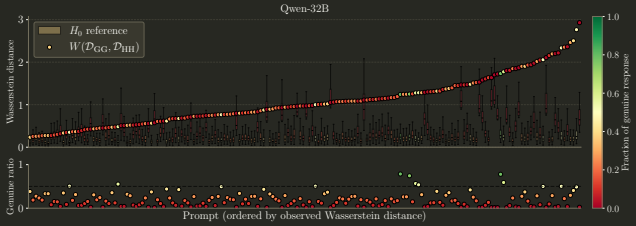
$$\mathbf{v} \propto (\mathbf{S}_W + \lambda \mathbf{I})^{-1}(\boldsymbol{\mu}_G - \boldsymbol{\mu}_H),$$

with a shared optimum $\lambda \approx 1.2$ across all models.



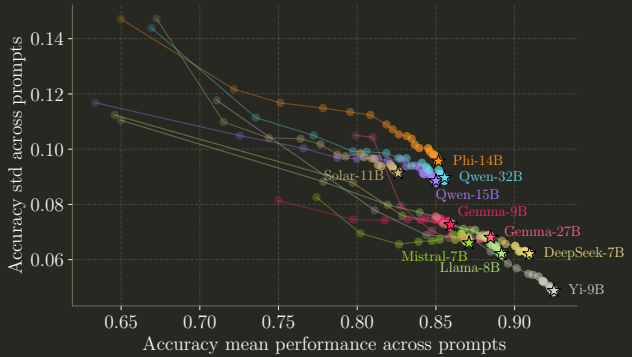
[*schematic*]

A.4 | Above chance, across the board



- 81.5%–94.5% of prompts separate above the null
- Fisher separability ratio: **1.13** → **7.26** (avg.)
- robust to encoder choice and to annotation noise

A.5 | APORIA-LP sample efficiency



- near-saturation by 30–50 labels
- avg. **F1 90.7%**, **acc 87.2%**
- performance independent of model size
- variance shrinks as labels grow

A.6 | External baselines on SOCRATES-300K

Method	Acc [%]	F1 [%]
INSIDE	71.0	74.9
Semantic Entropy	78.5	86.3
SelfCheckGPT (NLI)	82.0	89.8
APORIA-LP	**87.2**	**90.7**

[*Tasks differ (label propagation vs. detection); read as a proxy.*]

A.7 | Cross-dataset validation (CoQA)

- bridge dataset following INSIDE's protocol
- 3 models (OPT-6.7B, LLaMA-2 7B/13B)
- Fisher separability avg. **79.65** (vs. 7.26 on SOCRATES-300K)
- APORIA-LP: **F1** >**96%**, usable from N=10

A.8 | Why a Fisher projection?

Projection	Dim	Acc [%]	F1 [%]
Fisher	**1**	**86.9**	**90.5**
UMAP	3	~74	~75
whitened PCA	2–3	~74	~81
random proj.	15+	~71	~79

Takeaway

A single linear direction captures the discriminative signal
— interpretable, low-dimensional, and stable.