Simple Factuality Probes Detect Hallucinations in Long-Form Natural Language Generation

Jiatong Han, Neil Band, Tim G. J. Rudner, Yarin Gal (EMNLP 2025)

June 4, 2026

YunSeop Shin, Seoul National University, Statistics, IDEA LAB

Notation

- x: input prompt; $z \sim \pi_{\text{gen}}(\cdot \mid x)$: the long-form generation.
- π_{gen}: generator LM (produces the answer z).
- π_{aux}: auxiliary LM (decomposes z into atomic claims).
- π_{small}: optional *smaller* encoder LM; $\pi_{\text{enc}} \in \{\pi_{\text{gen}}, \pi_{\text{small}}\}$.
- c: a single **atomic claim**; $\mathcal{C}$: the set of claims extracted from z.
- $h_c = \text{GetHiddenState}(\pi_{\text{enc}}, c) \in \mathbb{R}^d$: hidden-state vector of claim c.
- f_{ret}: automatic fact-checker (e.g. Wikipedia / web search) that supplies the **label** $y_c = f_{\text{ret}}(c) \in \{0, 1\}$ — used only to label training data, *not* at inference.
- $\mathcal{D}_{\text{probe}} = \{(h_c, y_c)\}$: probe training set.
- $f(\cdot\,; \theta)$: the **Factuality Probe**; $\hat{p}_c = f(h_c; \theta) \in [0, 1]$.
- $S(c) = [s(c), e(c)]$: token span in z that implies claim c.

Motivation

- Frontier LLMs generate tens of thousands of tokens, yet still **hallucinate**.
- Goal: obtain **fine-grained factuality scores** — tell users *which spans are trustworthy* and which need checking.
- Existing methods are sampling-based (e.g. semantic entropy): they draw many generations per claim.

Key idea

An LLM's hidden states already encode factuality. Read it out with a lightweight probe, giving claim-level scores from a **single** generation call.

Preliminary: Atomic Facts

An **atomic fact** (the paper's **atomic claim** c) is a short statement carrying a single semantic piece of information.

How it is done. The auxiliary LM π_{aux} (GPT-4o-mini) decomposes z into atomic claims $\mathcal{C} = \{c_1, \ldots, c_m\}$ (FActScore; Min et al., 2023). the oracle f_{ret} then labels each c_i independently as $y_c \in \{0, 1\}$.

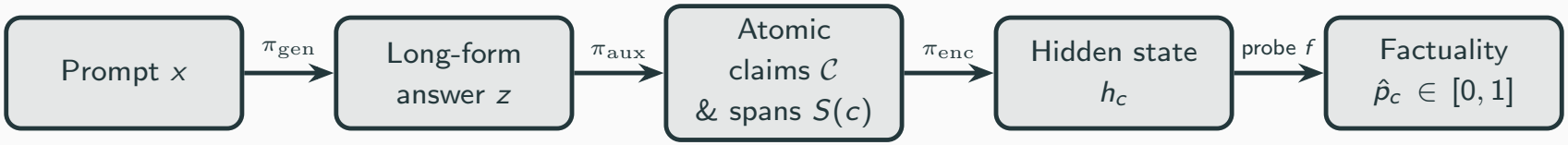
Example. *Input z:* "Marie Curie, a physicist, won the 1903 Nobel Prize in Physics and was the first woman to fly solo across the Atlantic." $\xrightarrow{\pi_{\text{aux}}}$

1. Marie Curie was a physicist. — **True**
2. Won the 1903 Nobel Prize in Physics. — **True**
3. First woman to fly solo across the Atlantic. — **False** (Amelia Earhart)

Core Idea & Pipeline

Central empirical finding

The hidden state of an atomic claim is **highly predictive** of whether that claim is *true* — recoverable by a simple linear map.



- **Span attribution** → a red/green *heatmap* over the whole generation.
- Only **one** generate → **no resampling**.

Training Factuality Probes

Stage 1 — build a supervised factuality dataset.

- For each prompt x: generate $z \sim \pi_{\text{gen}}(\cdot \mid x)$, then decompose with π_{aux} into atomic claims $\mathcal{C}$.
- Encode each claim $h_c = \text{GetHiddenState}(\pi_{\text{enc}}, c)$ and label it by the retrieval oracle $y_c = f_{\text{ret}}(c) \in \{0, 1\} \Rightarrow \mathcal{D}_{\text{probe}} = \{(h_c, y_c)\}$.

Stage 2 — fit the probe $f : h_c \mapsto p(y = \text{true} \mid c)$. Two simple choices:

1. **Sparse logistic regression** (L_1 penalty $\Rightarrow$ a few informative dims):

$$\min_{\theta} \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathcal{L}\big(y_c, f(h_c; \theta)\big) \; + \; \lambda \|\theta\|_1, \tag{1}$$

   with $\mathcal{L}$ the logistic loss and λ the sparsity weight.

2. **XGBoost**: gradient-boosted trees for nonlinear interactions.

Inference

Five steps to annotate a fresh generation:

1. **Generate.** $z \sim \pi_{\text{gen}}(\cdot \mid x)$.
2. **Decompose + attribute spans.** $(\mathcal{C}, \mathcal{S}) \sim \pi_{\text{aux}}$ with $S(c) = [s(c), e(c)]$; spans are **non-overlapping** (a shared token goes to the first claim detected).
3. **Extract hidden states.** $h_c = \text{GetHiddenState}(\pi_{\text{enc}}, c)$.
4. **Probe.** $\hat{p}_c = f(h_c) \in [0, 1]$, giving $\mathcal{R} = \{(c, \hat{p}_c) \mid c \in \mathcal{C}\}$.
5. **Visualize.** Heatmap each span $S(c)$ from red→green by $\hat{p}_c$.

Example

👤: Tell me a bio of Alexander Vovin.

🤖: … Vovin has also worked on the development of Ainu language teaching materials[F] and has been involved in various projects[T] to preserve[T] and promote Ainu culture.[F] He is currently a professor of Japanese and Korean languages and cultures[F] at the University of California, Davis.[F] Vovin's work has been recognized with several awards, including the Japan Foundation's Prize for Outstanding Contributions to Japanese Studies[T] in 2009.[T] He is a fellow of the American Academy of Arts and Sciences[?] ·...

🔬 Probe the atomic claim:
h = GetHiddenState(🤖, "Alexander Vovin was a fellow of the American Academy of Arts and Sciences.")
$\hat{p} = f(h) = 0.4838 \in [0,1]$

📊 Probability of factuality $\hat{p} = 0.4838$, Ground truth: False

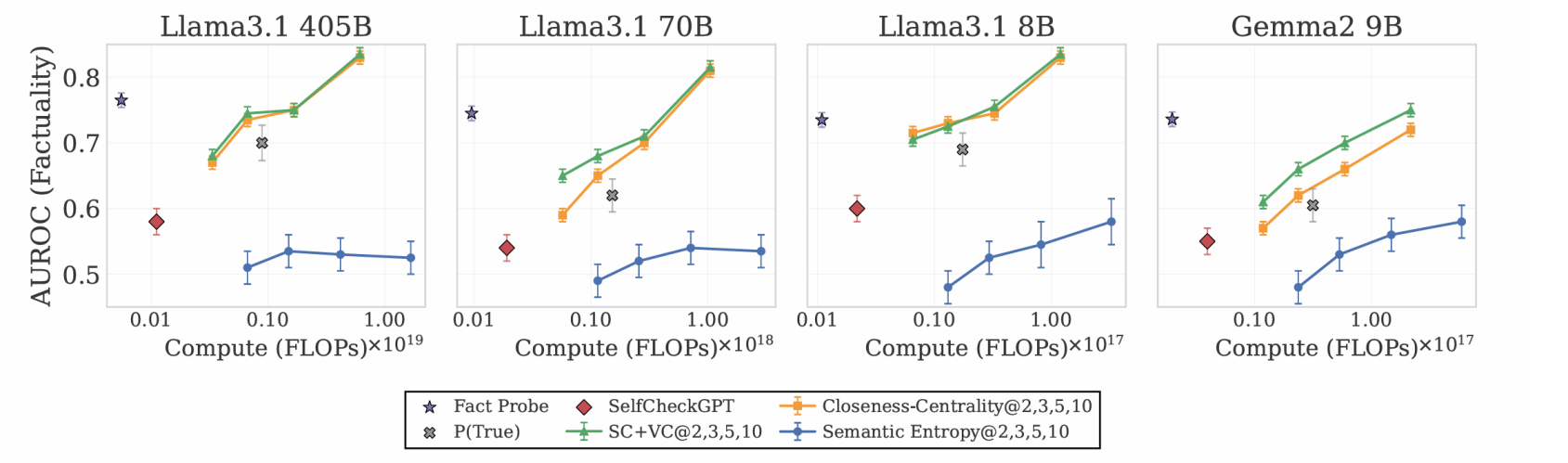
FActScore & Dataset

Two benchmarks supply the oracle labels f_{ret}:

- **FActScore** (Min et al., 2023): biographies; Wikipedia fact check.
- **LongFact** (Wei et al., 2024): diverse topics; web-search verification.

Domain-shift protocol: *train* on **LongFact**, *test* on **FActScore** labels (512-token cap; $\pi_{\mathrm{aux}} =$ GPT-4o-mini).

Experiments



π_{small} (f)	π_{gen}	**AUROC**
Llama 3.2 3B	Llama 3.2 3B	$0.7260_{\pm 0.0121}$
	Llama 3.1 8B	$0.7218_{\pm 0.0121}$
	Llama 3.1 70B	$0.7096_{\pm 0.0112}$
	Llama 3.1 405B	$0.6995_{\pm 0.0117}$
Llama 3.1 8B	Llama 3.1 8B	$0.7407_{\pm 0.0118}$
	Llama 3.1 70B	$0.7201_{\pm 0.0111}$
	Llama 3.1 405B	$0.7669_{\pm 0.0107}$
Llama 3.1 70B	Llama 3.1 70B	$0.7453_{\pm 0.0107}$
	Llama 3.1 405B	$0.7732_{\pm 0.0105}$