Are the Hidden States Hiding Something? Testing the Limits of Factuality-Encoding Capabilities in LLMs

Giovanni Servedio[1,2*] Alessandro De Bellis[1*] Dario Di Palma[1] Vito Walter Anelli[1] Tommaso Di Noia[1]
[1]Politecnico di Bari, Bari, Italy [2]Sapienza University of Rome, Italy
firstname.lastname@poliba.it

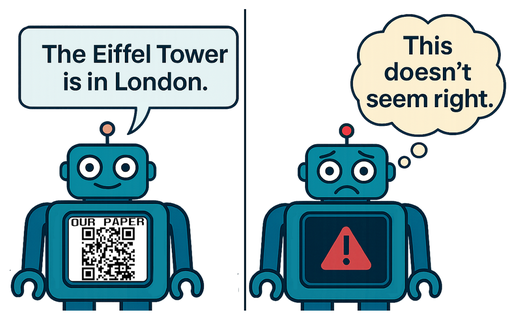
Can LLMs judge the factuality of their own generations?

Prior work (Azaria and Mitchell, 2023) suggests that LLM hidden activations contain a *truth signal*.
If robust, this would allow LLMs to *self-diagnose* the factuality **of their own generations** without relying on external verifiers.

Problem: Current approaches cannot properly assess **factuality self-evaluation** due to **unrealistic datasets** that do not adhere to the analyzed LLM generative process.

What We Propose:

1. The **reproduction** of the SAPLMA probe (Azaria and Mitchell, 2023) on *Llama-2-7B* and *OPT-6.7B*;
2. **Two new dataset construction pipelines:** A Perplexity-based Tabular Sampling and an LLM-based Fact Generation (QA datasets);
3. The **assessment of whether hidden states encode factuality signals** on LLM-aligned benchmarks.



Reproduction Study: *"The internal state of an LLM knows when it's lying?"* (Azaria and Mitchell, 2023)

- **True–False Dataset:** Fixed templates, object substitutions across topics;
- **SAPLMA:** 3-layer MLP (256→128→64);
- **Layer-wise probing:** 5 hidden layers (32, 28, 24, 20, 16);
- **Training protocol:** 5 training epochs, leave-one-topic-out evaluation strategy.

We train each model 20 times to replicate results.

- **Reproducibility:** Results closely match the original study.
- **Performance:** Middle layers (16–24) outperform final layers.
- **Thresholding:** Optimal thresholding boosts accuracy.

Conclusion: SAPLMA is reproducible. Hidden states, especially in middle layers, retain factuality signals.

LLM-aligned dataset construction as a step toward more reliable self-evaluation

Goal: Build **LLM-aligned datasets** to test an LLM's ability to judge the truth of its *own* outputs, an assessment that is unreliable with externally generated True–False sets.

Limitations of the Original True–False Dataset (Azaria and Mitchell, 2023):

- **Template rigidity.** Fixed patterns curb linguistic diversity and overfit the probe;
- **Distribution misalignment.** Statements are not drawn from LLM distribution;
- **Knowledge mismatch.** LLMs cannot judge facts they never saw;
- **Varying cardinality.** Statements have different admissible object substitutions.

Proposed Remedies

1. Perplexity-based Tabular Sampling

Goal: Improve false statement quality in synthetic datasets via perplexity-guided sampling.

1. Insert correct entity-property pairs into sentence templates;
2. Replace the property to generate false candidates; compute LLM perplexity;
3. Keep examples where the true statement has low perplexity (top-k, $k = \alpha|C|$);
4. Retain false candidates whose perplexity is within $(1+\beta)$ of the true's;
5. Compute scores from perplexity (lower = better); normalize into probabilities;
6. Apply top-k and nucleus sampling to select the final false statement.

2. QA-Based Fact Generation

Goal: Build a realistic, diverse true/false factual dataset using LLM-generated statements, addressing the limitations of template-based generation.

1. Begin with a QA dataset $D_{QA} = \{(q_i, a_i)\}$;
2. For each question q_i, prompt the LLM K times to generate answers $\{a_{i,k}^M\}$;
3. An oracle LLM assigns veracity labels $v_{i,k}^M \in \{0,1\}$, conditioned on q_i and a_i;
4. Compute the ratio of correct answers $p_i = mean(v_{i,k}^M)$;
5. Select questions with mixed responses where $|p_i - 0.5| < \tau$;
6. Collect the answers and labels that meet this filter: $D_{\text{Facts}} = \{(a_{i,k}^M, v_{i,k}^M)\}$.

Outcome: Datasets that better reflect the LLM's learned distribution and enable more reliable evaluation of its self-assessment abilities.

LLMs show signal of factuality, yet struggle on LLM-generated data

We test SAPLMA probe classifiers on two new datasets:

1. **Refined True-False dataset** with perplexity-based sampling

Layer	Training Data	Cities		Inventions		Elements		Animals		Companies		Average	
		Llama 2-7b	OPT-6.7b	Llama 2-7b	OPT-6.7b	Llama 2-7b	OPT-6.7b	Llama 2-7b	OPT-6.7b	Llama 2-7b	OPT-6.7b	Llama 2-7b	OPT-6.7b
last	Orig.	0.6882	0.5724	0.6409	0.5094	0.6314	0.5482	0.5685	0.5259	0.6290	0.6984	0.6316	0.5709
	Novel	0.6365	0.6143	0.6101	0.5144	0.5623	0.5293	0.5461	0.4846	0.7414	0.7211	0.6311	0.5720
28	Orig.	0.7056	0.5870	0.7001	0.5178	0.6161	0.5832	0.6013	**0.5566**	0.7079	0.7234	0.6662	0.5936
	Novel	0.5091	0.6057	0.6591	0.5473	0.5665	0.5655	0.6052	0.4662	0.7061	0.7065	0.6092	0.5811
24	Orig.	0.8286	0.7026	0.7250	0.6022	0.6432	0.5918	0.6310	0.5439	0.7121	0.7366	0.7080	0.6354
	Novel	0.6025	0.6710	0.6609	0.6248	0.5763	0.5993	0.5836	0.4868	0.7868	0.7366	0.6420	0.6225
20	Orig.	0.8272	0.7313	0.7741	**0.6230**	0.6492	**0.6255**	0.5832	0.5075	0.7583	0.7502	0.7184	**0.6475**
	Novel	0.7382	**0.7528**	0.6973	0.6131	0.6051	0.5986	0.6190	0.4807	**0.8270**	**0.7566**	0.6973	0.6404
16	Orig.	0.8941	0.6433	0.7888	0.5698	**0.6801**	0.5930	0.5836	0.3816	0.7768	0.7426	0.7447	0.5860
	Novel	**0.9301**	0.7505	**0.7961**	0.5644	0.6623	0.5568	**0.6319**	0.5307	0.8265	0.7231	**0.7694**	0.6151

Table 1. Accuracy values obtained training SAPLMA on the original True-False dataset and on our refined version, then **tested on our refined version**. 'Orig.' denotes the 'original True-False dataset as training data, while 'Novel' denotes our version of the True-False dataset as training data.

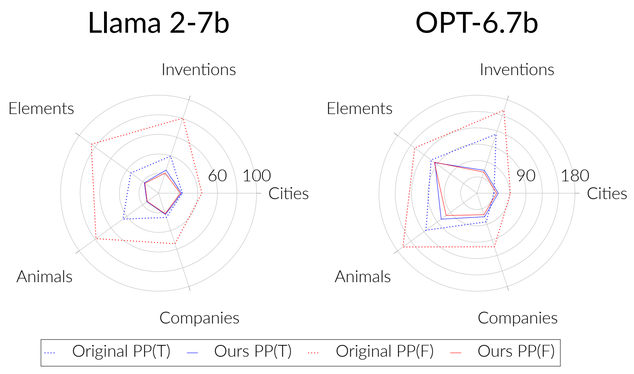


Figure 1. Comparison of average perplexity scores (True vs False sentences) for the Original dataset by Azaria and Mitchell (2023) and our refined version.

- Results support Azaria and Mitchell (2023), **factuality is encoded in LLM hidden states**.
- Probes remained robust even on True-False sentences with **similar perplexities**.
- This result supports the idea that **hidden states reveal deeper information** than what is apparent at the surface level.

2. **LLM-generated dataset** obtained by sampling answers from TriviaQA.

Dataset		Threshold = 0.5				
		last	28	24	20	16
billturnbull	Llama	.561	.576	.618	.621	.628
	OPT	.547	.558	.591	.551	.537
derby*	Llama	.568	.581	.575	.596	.617
	OPT	.553	.564	.584	.587	.572
quiz4free	Llama	.564	.547	.523	.561	.589
	OPT	.559	.559	.575	.581	.560
quizguy	Llama	.578	.585	.588	.607	.635
	OPT	.579	.583	.589	.590	.584
triviabug	Llama	.494	.518	.521	.525	.538
	OPT	.620	.624	.607	.596	.528
businessballs	Llama	.566	.558	.565	.574	.582
	OPT	.559	.558	.578	.570	.553
jetpunk	Llama	.587	.627	.620	.643	.654
	OPT	.606	.612	.614	.596	.621
odquiz	Llama	.551	.536	.546	.562	.573
	OPT	.560	.573	.583	.583	.542
quiz-zone	Llama	.556	.557	.558	.565	.611
	OPT	.569	.570	.582	.592	.552
quizballs	Llama	.603	.575	.572	.578	.571
	OPT	.558	.565	.574	.571	.540
quizwise	Llama	.560	.565	.579	.609	.618
	OPT	.560	.563	.565	.577	.540
sfquiz	Llama	.568	.554	.554	.559	.575
	OPT	.584	.591	.589	.590	.547
triviacountry	Llama	.536	.554	.559	.556	.566
	OPT	.536	.551	.550	.587	.534
wrexham**	Llama	.570	.563	.569	.553	.565
	OPT	.548	.573	.578	.584	.554
Average	Llama	.562	.564	.568	.579	.594
	OPT	.567	.575	.583	.583	.555

*: derby is adopted as abbreviation of derbyshirepubquizleague.
**: wrexham is adopted as abbreviation of wrexhamquizleague.

Table 2. Performance of SAPLMA on a fact dataset generated from TriviaQA. The original topic-wise leave-one-out strategy is adopted.

Experimental Setup:

- Evaluated SAPLMA classifier on LLM-generated sentences derived from TriviaQA.
- Classifiers trained and tested with a leave-one-out strategy.

Key Findings:

- **Low Accuracy:** Current probes perform poorly in assessing factuality on LLM-generated data.
- **Threshold Optimization Fails:** Tuning the classification threshold does not significantly improve results.
- **Consistency Across Datasets:** Similar underperformance observed on SQuAD 2.0 and TruthfulQA.

Conclusion:

- Probes trained on the True-False dataset do *not* generalize well to LLM-generated facts.
- TriviaQA's open-domain questions likely introduce more nuanced and complex facts than prior datasets.

Conclusion: Factuality appears encoded in LLMs on synthetic data, but fails to generalize to LLM-generated content, revealing the need for more reliable probes.

References

Amos Azaria and Tom M. Mitchell. The internal state of an LLM knows when it's lying. In *EMNLP (Findings)*, pages 967–976. Association for Computational Linguistics, 2023.