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, Italy [2]Sapienza University of Rome, Italy
name.surname@poliba.it

Abstract

Factual hallucinations are a major challenge for Large Language Models (LLMs). They undermine reliability and user trust by generating inaccurate or fabricated content. Recent studies suggest that when generating false statements, the internal states of LLMs encode information about truthfulness. However, these studies often rely on synthetic datasets that lack realism, which limits generalization when evaluating the factual accuracy of text generated by the model itself. In this paper, we challenge the findings of previous work by investigating truthfulness encoding capabilities, leading to the generation of a more realistic and challenging dataset. Specifically, we extend previous work by introducing: (1) a strategy for sampling plausible true-false factoid sentences from tabular data and (2) a procedure for generating realistic, LLM-dependent true-false datasets from Question Answering collections. Our analysis of two open-source LLMs reveals that while the findings from previous studies are partially validated, generalization to LLM-generated datasets remains challenging. This study provides a foundation for future research on factuality in LLMs and offers practical guidelines for more effective evaluation. Code is provided at our `GitHub Repository`.

1 Introduction

In the last few years, Large Language Models (LLMs) have shown outstanding abilities in natural language processing tasks and beyond (Biancofiore et al., 2025; Di Palma, 2023). Nevertheless, factual hallucinations (Zhang et al., 2023) represent a significant obstacle, limiting their reliability and hindering their safe deployment in real-world applications (Di Palma et al., 2025) such as healthcare (Pham and Vo, 2024), education (Upadhyay et al., 2023), legal advice (Dahl et al., 2024), and language understanding (De Bellis et al., 2024; Anelli et al., 2022). Hallucinations occur when an LLM generates content that is syntactically coherent but factually inaccurate, decreasing trust in AI systems (Huang et al., 2024). Recent research suggests that LLMs may encode internal representations of factuality in their hidden states, indicating an awareness of whether a generated statement is true or false (Chen et al., 2024). These efforts led to the development of approaches to evaluate the factual accuracy of the LLM outputs given their internal representations (*factuality "self-evaluation"*). Self-evaluation can be used to identify gaps in the knowledge of an LLM, improving truthfulness and transparency through abstention mechanisms (Feng et al., 2024), fact verification (Wadden et al., 2020), and self-correction (Ji et al., 2023). Azaria and Mitchell (2023) suggest that LLMs have "some internal notion as to whether a sentence is true or false, as this information is required for generating (or predicting) following tokens." Based on this assumption, they propose a neural classifier to discern factual from non-factual statements based on hidden layer activations. However, the datasets used to evaluate the probe present limitations since they contain trivially incorrect statements (e.g., "The zebra uses flying for locomotion") that easily fail to align with the generative patterns of LLMs. Additionally, the false statements are generated using random substitutions of the true terms with little regard for the plausibility of negative samples. This misalignment not only weakens the generalizability of results but also raises concerns about the applicability of these models to real-world scenarios where false statements may be subtle or nuanced. This study addresses these gaps by generating more plausible datasets (see Figure 1) to explore LLM factuality encoding and evaluating refined models. The primary contributions are:

1. We **reproduce** the methodology of Azaria and Mitchell (2023) to ensure transparency.

[†]Corresponding authors.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6089–6104
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

2. We **propose two strategies to generate realistic datasets** and discuss how well the original and newly developed models generalize.

Specifically, we extend their work with two novel dataset creation strategies and design a strategy that better fits the factuality self-evaluation task, introducing:

- A **perplexity-based negative sampling strategy** that enhances the original generation mechanism and leverages the LLM token distribution.
- A **novel strategy to sample realistic LLM-generated facts**, leveraging Question Answering datasets to elicit responses from the LLM.

Through this analysis, we lay the groundwork for more robust factuality assessments and offer practical guidelines for enhancing the reliability of LLMs in diverse applications.

2 Reproduction of Prior Work: Settings

In their study, Azaria and Mitchell (2023) investigate whether LLMs internally represent the factuality of sentences. This section summarizes the dataset generation approach they employed and the specific probing architecture used in their study.

2.1 Dataset Generation Strategy

To explore whether LLMs internally represent the factuality of statements, the authors constructed a 'True-False' dataset of facts labeled as either True or False, covering six disjoint topics: *Cities*, *Inventions*, *Chemical Elements*, *Animals*, *Companies*, and *Scientific Facts*. To generate the dataset, for the first five topics, i.e. with the exception of *Scientific Facts*, the authors selected **true statements** from reliable sources (see Table 7 in the appendix) and produced **false statements**, replacing part of a true statement (e.g. *"Hydrogen has an atomic number of 1"*) with randomly sampled incorrect information (*"Hydrogen has an atomic number of 34"*). Meanwhile, for *Scientific Facts*, they employed ChatGPT (13 Feb 2023) as a generator of true and false sentences, and two human annotators manually verified their correctness. The authors publicly release the dataset, which we refer to as the *"True-False dataset"*. Furthermore, the authors constructed a second dataset using the OPT-6.7b model, which we refer to as the '**OPT-Generated Dataset**'. To create this dataset, the model was prompted with a true statement absent from the True-False dataset and then used to generate a subsequent sentence. The responses were manually fact-checked and annotated by three independent human judges. Non-factual responses were filtered out, resulting in a final set of 245 statements.

2.2 Internal States Analysis via SAPLMA

To investigate whether LLMs internally represent the factuality of statements, Azaria and Mitchell (2023) developed a probe (**S**tatement **A**ccuracy **P**rediction based on **L**anguage **M**odel **A**ctivations) that predicts the factual accuracy of a statement by analyzing the hidden layer activations of an LLM. SAPLMA is a feedforward neural network designed to classify statements as true or false. It consists of three hidden layers (256, 128, 64) and a sigmoid output activation. The model is trained using the Adam optimizer for five epochs without hyperparameter tuning. The authors studied two LLMs, namely OPT-6.7b (Zhang et al., 2022) and Llama 2-7b (Touvron et al., 2023), both consisting of 32 layers. To identify which layers best capture factuality, they trained five SAPLMA models, forwarding each statement in the True-False Dataset as input to the LLMs and extracting the corresponding activation values from the 32nd (last layer), 28th, 24th, 20th, and 16th layers. These activations serve as input for training the classifiers. To ensure generalizability, i.e., making SAPLMA independent of specific topics, the authors adopted a cross-validation strategy using a leave-one-topic-out approach to train the classifier on five topics and test the probe on the held-out topic.

2.3 Reproducibility Settings

In this section, we provide details on the datasets and experimental settings for reproducing the work of Azaria and Mitchell (2023). Our goal is to answer the Research Question (**RQ0**): *Can we reproduce the results reported by Azaria and Mitchell (2023)?"* Although the code is not publicly accessible, the authors made it available upon request.

Dataset generation. The code provided by Azaria and Mitchell (2023) contains all the necessary material to recreate the entire dataset generation process. However, their template-matching code is influenced by randomness in the generation of false statements, and a random seed is not set. Due to this non-deterministic behavior, recreating their dataset using the original code was unfeasible. However, the authors released their dataset, allowing us to reproduce their exact dataset settings.

SAPLMA reproducibility. To reproduce the results of the original study, we trained 20 SAPLMA

Layer		Cities		Inventions		Elements		Animals		Companies		Facts		Average	
		Llama2	OPT-6.7b	Llama2	OPT-6.7b	Llama2	OPT-6.7b	Llama2	OPT-6.7b	Llama2	OPT-6.7b	Llama2	OPT-6.7b	Llama2	OPT-6.7b
Last	Orig.	0.7574	0.7796	0.6735	0.5696	0.6814	0.5760	0.7338	0.6022	0.6736	0.6925	0.7444	0.6498	0.7107	0.6449
	Repr.	0.7939	0.7836	0.7470	0.5603	0.7057	0.5656	0.7133	0.5984	0.6463	0.6900	0.7894	0.6640	0.7326	0.6437
28	Orig.	0.8146	0.7732	0.7207	0.5761	0.6767	0.5907	0.7249	0.5777	0.6894	0.7247	0.7662	0.6618	0.7321	0.6507
	Repr.	0.8261	0.8014	0.7221	0.5938	0.6746	0.5931	0.7046	0.5945	0.6860	0.7252	0.7976	0.6639	0.7351	0.6620
24	Orig.	0.8722	0.7963	0.7816	0.6712	0.6849	0.6211	0.7394	0.5800	0.7094	0.7758	0.7858	0.6868	0.7622	0.6886
	Repr.	0.8619	0.8043	0.7737	0.6604	0.6789	0.6172	0.7415	0.6095	0.7049	0.7844	0.7910	0.6804	0.7586	0.6927
20	Orig.	0.8820	**0.8125**	0.8459	**0.7268**	0.6950	0.6197	0.7758	0.6058	0.8319	0.8122	0.8053	**0.6819**	0.8060	0.7098
	Repr.	0.8672	0.8118	0.8584	0.7222	0.6761	**0.6218**	0.7736	**0.6208**	0.8254	**0.8160**	0.8065	0.6734	0.8012	**0.7110**
16	Orig.	**0.9223**	0.7435	**0.8938**	0.6400	0.6939	0.5645	0.7774	0.5800	0.8658	0.7570	**0.8254**	0.6237	**0.8298**	0.6515
	Repr.	0.9174	0.7554	0.8847	0.6403	**0.7005**	0.5732	**0.7883**	0.5693	**0.8672**	0.7760	0.8104	0.6340	0.8281	0.6580

Table 1: Replicated SAPLMA performance on the True-False dataset across the selected layers. The results labeled as ‘Orig.’ are taken from the original work, while those labeled as ‘Repr.’ are the replicated results from this study.

probes for each of the following layers: the 32nd, 28th, 24th, 20th, and 16th, over 5 epochs, resulting in a total of 100 probes. We employed Llama 2-7b and OPT-6.7b, both using half-precision (16-bit float) parameters, with a default temperature of 0.8 for Llama 2-7b and 1.0 for OPT-6.7b. The hardware used for the experiments was an Intel(R) Core(TM) i7-5820K paired with an NVIDIA RTX 3090 graphics card. The authors do not specify from which token they extract the associated hidden state. However, code inspection led to the identification of the last token as the target state.

Model		Cities	Inventions	Elements	Animals	Companies	Facts	Average
BERT	Orig.	0.5357	0.5537	0.5645	0.5228	0.5533	0.5302	0.5434
	Repr.	0.5257	0.5611	0.5435	0.5603	0.5302	0.5361	0.5428
3-shot	Orig.	0.5410	0.4799	0.5685	0.5650	0.5538	0.5164	0.5374
	Repr.	0.5416	0.4800	0.5685	0.5652	0.5539	0.5115	0.5368
5-shot	Orig.	0.5416	0.4799	0.5676	0.5643	0.5540	0.5148	0.5370
	Repr.	0.5416	0.4800	0.5676	0.5643	0.5540	0.5082	0.5359
It-is-true	Orig.	0.5230	0.5068	0.5688	0.4851	0.6883	0.5840	0.5593
	Repr.	0.5233	0.5046	0.5688	0.4831	0.6875	0.5856	0.5588

Table 2: Replicated baselines performance on the True-False dataset. The results labeled as ‘Orig.’ are taken from the original work, while those labeled as ‘Repr.’ are the replicated results from this study.

3 Experimental Reproducibility Results

To answer **RQ0**, we report the results for the reproduction of the experiments in Tables 1, 2 and 3. The results labeled with ‘Orig.’ are retrieved from the original work, while the ones achieved in the reproducibility study are labeled with ‘Repr.’.

3.1 Reproduction of SAPLMA Results on the True-False Dataset

Table 1 reports SAPLMA’s performance on the True-False dataset across different layers of Llama 2-7b and OPT-6.7b for six categories: Cities, Inventions, Elements, Animals, Companies, and Facts. Results indicate that Llama 2-7b consistently outperforms OPT-6.7b across all layers and categories.

Layer		Accuracy	AUC	Accuracy with Optimal Threshold	Average Threshold
Last-layer	Orig.	0.6187	0.7587	0.7052	0.8687
	Repr.	0.6406	0.7720	0.7264	0.8910
28th-layer	Orig.	0.6362	0.7614	0.7134	0.8838
	Repr.	0.6410	0.7686	0.7203	0.8276
24th-layer	Orig.	0.6134	0.7435	0.6988	0.8801
	Repr.	0.6206	0.7496	0.6973	0.8500
20th-layer	Orig.	0.6029	0.7182	0.6587	0.9063
	Repr.	0.5965	0.7183	0.6669	0.8868
Middle-layer	Orig.	0.5566	0.6610	0.6500	0.8123
	Repr.	0.5579	0.6760	0.6468	0.7948
BERT	Orig.	0.5115	0.5989	0.5705	0.9403
	Repr.	0.5522	0.6092	0.5689	0.7939
3-shot	Orig.	0.5041	0.4845	-	-
	Repr.	0.5041	0.4845	-	-
5-shot	Orig.	0.5125	0.4822	-	-
	Repr.	0.5125	0.4822	-	-

Table 3: Reproduced SAPLMA performance on the **OPT-Generated Dataset** (Section 2.1). The results labeled as ‘Orig.’ are taken from the original work, while those labeled as ‘Repr.’ are the reproduced results.

Middle layers (16, 20, 24) achieve the highest performance, while accuracy declines toward the final layer. The reproduced results closely align with the original findings, with minor deviations observed across specific categories and layers. Moreover, although OPT-6.7b shows greater variability, the ranking of layers remains consistent with the original work. Both experiments confirm that factuality information is more effectively encoded in the middle layers (16–24) than in the final layer. Additionally, we reproduce the baseline used by the authors to compare SAPLMA. Specifically, their baseline includes a trained SAPLMA on BERT activations, a few-shot approach where the LLM is prompted with a sentence and asked to label it as ‘True’ or ‘False,’ and an ‘It-is-true’ test. In this test, the LLMs were asked: *Is it true* that X? and *Is it false that X?*, where X is a dataset sample. A response was considered correct if the model assigned a higher probability to the ‘True’ token. Table 2 summarizes the baseline performance. BERT achieves the highest average performance among non-prompted methods, demonstrating its effectiveness in factual classification, and the results indicate high reproducibility. The It-is-true baseline yields the highest average performance, particularly in the 'Companies' topic. The results demonstrate consistent reproducibility.

In general, the reproduction on the True-False Dataset is considered successful based on the following observations: (i) the overall performance trends remain consistent, with Llama 2-7b outperforming OPT-6.7b; (ii) observed deviations are minor and do not indicate fundamental inconsistencies; (iii) the relative ranking of layers remains unchanged, reinforcing previous findings; and (iv) baseline methods retain their rankings, confirming the validity of the original results.

3.2 Reproduction of the Results on the OPT-Generated Dataset

Table 3 presents the results of the reproduced experiments on the OPT-Generated Dataset. Performance is evaluated using Accuracy, AUC, Accuracy with an Optimal Threshold (selected by estimating it from a held-out validation set), and the Average Optimal Threshold. The final layers, specifically the 28th and final layers, outperform the middle and lower layers in terms of Accuracy and AUC. The reproduced results closely align with the original findings, exhibiting only minor variations. Accuracy with anoptimal threshold consistently exceeds raw accuracy, suggesting that tuning the decision boundary improves performance.

Regarding baselines, BERT exhibits lower accuracy compared to the LLMs' last layers, with slight improvements over the original results. Notably, the 3- and 5-shot prompting results were identical between the original and reproduced experiments.

RQ0: *Can we reproduce the results reported by Azaria and Mitchell (2023)?*

This reproducibility study demonstrates a high degree of alignment with the original results, confirming the validity of previous findings. It shows that the ranking and trends remain unchanged, reinforcing the robustness of the results.

4 Novel Dataset Generation Strategies for Factuality Self-Evaluation

This section introduces two novel dataset generation strategies to investigate LLM factuality self-evaluation. To contextualize our approach, we first examine the limitations of the True-False dataset by Azaria and Mitchell (2023). While it provides a structured framework for evaluating LLMs, its construction imposes constraints that may limit generalizability of the findings. Specifically, true and false statements are derived from tabular data using predefined templates. We argue that this approach suffers from several limitations:

- **Adherence to predefined templates:** The use of fixed templates limits the linguistic expressiveness of the dataset, potentially constraining the probe classifier's ability to generalize beyond rigidly structured statements (e.g., <*company*> *operates in the industry of* <*industry*>).
- **Distribution misalignment:** The statements are constructed from tabular data rather than generated by the LLM itself. As a result, the dataset may not align with the LLM's generative distribution. For instance, a niche true fact in the dataset –but unknown by the model– could have high perplexity for the LLM, undermining the study's core premise: evaluating an LLM's intrinsic ability to "judge" its false claims.
- **Lack of consideration for the LLM's knowledge state:** LLMs exhibit uneven factual knowledge based on their pretraining data, with strengths in some domains and gaps in others. A model can assess a statement factuality only if it has prior exposure to it. The dataset does not account for these inconsistencies: it evaluates whether an LLM can detect factual errors without considering whether the model actually possesses knowledge of the fact.
- **Differences in cardinality:** Some properties in the dataset have fewer admissible values, making certain facts easier to evaluate. For example, a statement like *"<element> appears in its standard state as _"* has fewer possible values (i.e.,*{Solid, Liquid, Gas}*) compared to statements like *"<city> is a city in _"* that involve a broader set. This imbalance in complexity may bias the evaluation process.

These limitations impact the interpretability of results when evaluating an LLM's internal representations of factuality. To ensure a more realistic assessment of an LLM's self-evaluation of factuality, we propose two strategies to address these issues. The first strategy samples statements from tabular data to better align with the LLM's generative predictions. The second strategy involves sampling LLM-generated facts as answers to questions from a well-known Question Answering dataset.

4.1 Perplexity-based Dataset Construction

This section presents a novel dataset generation strategy to address limitations in the Azaria and Mitchell (2023) True-False dataset, particularly distribution misalignment and implausible negative samples. To improve negative sampling based on random property-object substitutions, we introduce a **perplexity-guided probabilistic sampling** method, which re-weights false statements based on perplexity for better alignment with LLM output distributions. Since perplexity depends on the model, the same LLM under evaluation must generate false statements to ensure consistency, resulting in a model-dependent dataset tailored to the characteristics and biases of the LLM being studied.

True statements are initially constructed, as in Azaria and Mitchell (2023), by directly inserting correct entity-property pairs into pre-defined sentence templates. Regarding the false statements generation, we proceed as follows: (i) for each true statement, all unique alternative property values are gathered from the entire dataset; (ii) a candidate sentence is created for each alternative property by inserting it into the corresponding template; (iii) the target LLM (i.e., OPT-6.7b or Llama 2-7b) computes the perplexity of each candidate sentence. This perplexity score serves as a plausibility metric, with lower perplexity indicating a more plausible (yet incorrect) statement.

Given a true statement, we define C as the set of potential candidate sentences, which includes the true statement. Furthermore, $C' \subset C$ is defined as the subset of candidate false sentences (i.e. excluding the true statement). Candidates $c \in C$ are ranked based on their perplexity scores, with lower scores indicating higher plausibility.

Since perplexity can be interpreted as a measure of plausibility, we operate under the assumption that an LLM possesses factual knowledge about a fact if the fact is assigned a "sufficiently low perplexity". Conversely, a high perplexity score for the true statement suggests that the LLM lacks knowledge of the fact. Given the limitations discussed earlier, we aim to evaluate the LLM ability to discern between true and false statements when it possesses the relevant knowledge. Therefore, we exclude instances where the LLM exhibits limited knowledge about the true statement: if the true statement does not rank among the lowest k perplexity candidates, the generation of that instance is discontinued, and the next true fact is considered. In practice, we define k as $k = \alpha|C|$, where α is a hyperparameter ($0 < \alpha < 1$). This accounts for the varying cardinality of the property ranges in the dataset, ensuring that the threshold for "sufficiently low perplexity" is adjusted based on the number of possible values for a given property.

In addition, we want to simulate a real hallucination scenario where the LLM is uncertain between the true fact and plausible alternatives: given the perplexity score function $PP(\cdot)$, all false candidates c with a perplexity score $PP(c) < (1+\beta)PP(true)$, where $0 < \beta < 1$ is a hyperparameter, are considered, resulting in a reduced set of candidates $C^{\star}$. A min-max normalization is applied to their perplexity,

$$\text{NormPP}(c) = \frac{\text{PP}(c) - \min_{c \in C} \text{PP}(c)}{\max_{c \in C} \text{PP}(c) - \min_{c \in C} \text{PP}(c)}, \tag{1}$$

The normalized perplexities are transformed using a plausibility score function $s(\cdot)$, i.e., lower perplexity scores result in higher plausibility scores. The scores are then normalized to ensure that they sum to 1 and are treated as a probability distribution over the candidates:

$$s(c) = e^{-\text{NormPP}(c)}, \quad P(c_i) = \frac{s(c_i)}{\sum_{c_j \in C^{\star}} s(c_j)}. \tag{2}$$

The normalization guarantees that $P(c_i)$ values are suitable for sampling. Finally, a mixture of top-k and nucleus sampling (Holtzman et al., 2020) is employed to sample the candidate for insertion into the template. Specifically, we apply top-k and nucleus sampling sequentially: we select the top-k highest-scoring candidates and then refine this set by choosing the smallest subset whose cumulative probability reaches a predefined threshold. This process generates a coherent yet factually incorrect statement that is more realistic and closely aligned with the LLM internal token prediction patterns. In Section 5.1 we detail the selected values for the hyperparameters α, β, k, and p.

4.2 LLM-Generated Dataset Construction

The strategy in Section 4.1 constructs a balanced true-false dataset from tabular data but has inherent limitations. While enabling direct comparison with Azaria and Mitchell (2023), it restricts generative models to rigid templates, limiting expressiveness. Additionally, its reliance on fixed candidate sets can lead to easily classifiable false statements when the set is small. Finally, template-based sampling reduces diversity, likely due to token bias: we found LMs consistently assigning lower perplexity to certain property-object pairs, regardless of the subject. To more realistically assess an LLM factuality self-evaluation, we propose generating both true and false statements directly from the model, overcoming the limitations of template-based approaches. This involves (a) a method to elicit diverse factoid statements from the LLM, and (b) a strategy to annotate statement veracity, addressing biases and inconsistencies in the process. Consider a Question Answering dataset composed of N questions, $D_{QA} = \{(q_i, a_i)\}_{i=1}^{N}$, where each question q_i has a corresponding ground-truth answer a_i. Given a LLM $\mathcal{M}$, we prompt it with each question K times, yielding a set of generated answers. This results in an extended dataset $D_{QA}^{\mathcal{M}} = \{(q_i, a_i, \{a_{i,k}^{\mathcal{M}}\}_{k=1}^{K})\}_{i=1}^{N}$. Following LLM-as-judge (Gu et al., 2024; Calderon et al., 2025) practices, the LLM-generated answers in $D_{QA}^{\mathcal{M}}$ can be annotated using an oracle LLM, which we assume is able to evaluate the veracity of each answer $a_{i,k}^{\mathcal{M}}$ given the ground-truth answer a_i and the question q_i. This operation results in an annotated dataset

$$\hat{D}_{QA}^{\mathcal{M}} = \{(q_i, a_i, \{a_{i,k}^{\mathcal{M}}, v_{i,k}^{\mathcal{M}}\}_{k=1}^{K})\}_{i=1}^{N}, \quad (3)$$

where $v_{i,k}$ is a veracity label assigned by the oracle, indicating whether the generated answer $\hat{a}_{i,k}$ is correct ($v_{i,k} = 1$) or incorrect ($v_{i,k} = 0$).

Increasing K enhances the reliability of responses by offering a more accurate evaluation of the LLM knowledge state regarding a question. This evaluation can lead to three possible outcomes: (i) a high proportion of correct answers suggests the LLM fully understands the facts; (ii) a high proportion of incorrect answers indicates the LLM lacks or has partial knowledge, preventing correct responses; (iii) a mix of correct and incorrect answers implies knowledge with a tendency toward hallucination. This study assumes that an LLM can only encode factuality regarding a generated fact if it has some knowledge about it. Therefore, we focus exclusively on the third scenario. This scenario also naturally produces a balanced dataset, including true and false variations of the same fact. We define the correct answer ratio $p_i^{\mathcal{M}} = \frac{1}{K}\sum_{k=1}^{K} v_{i,k}^{\mathcal{M}}$. We consider questions whose p_i is around 0.5 with a tolerance hyperparameter τ, that is $|p_i^{\mathcal{M}} - 0.5| < \tau$. The dataset is obtained by selecting the answers

Dataset	Llama 2-7b		OPT-6.7b	
	Sentences	(%) True	Sentences	(%) True
Cities	674	50	756	50
Inventions	336	50	202	50
Elements	118	50	220	50
Animals	116	50	114	50
Companies	326	50	310	50

Table 4: Novel generation of our True-False dataset, following the approach described in Section 4.1, including number of sentences and percentage of true samples.

and their veracity labels satisfying the condition:

$$D_{Facts}^{\mathcal{M}} = \{(a_{i,k}^{\mathcal{M}}, v_{i,k}^{\mathcal{M}}) : |p_i^{\mathcal{M}} - 0.5| < \tau\}_{i=1}^{N}. \quad (4)$$

5 Experiments and Discussion

This section describes the experimental setup used to extend the prior investigation, presents the results, and discusses their implications. We address the following Research Questions:

RQ1 *Are probing classifiers trained as in Azaria and Mitchell (2023) capable of generalizing to True-False sentences with similar perplexity?*

RQ2 *Can the same probes generalize to facts generated by LLMs?*

5.1 Dataset Generation Setup

To generate the datasets for training the probe and alleviate the limitations of Azaria and Mitchell (2023), we employ the perplexity-based sampling procedure described in Section 4.1. Specifically, we set the needed hyperparameters as follows: $\alpha = 0.1$, $\beta = 0.1$, $k = 10$ and $p = 0.9$. We excluded the *Scientific Facts* topic from our perplexity-based sampling procedure, as its original version was generated by ChatGPT and not from tabular data (additional details in Appendix B). Table 4 summarizes the statistics for the two refined True-False datasets, while Figure 1 shows the differences in average perplexity between the original and the proposed datasets.

We construct an LLM-generated dataset following the procedure described in Section 4.2. Our experiments use TriviaQA (Joshi et al., 2017), a dataset of question-answer pairs collected from 14 trivia and quiz-league websites. Given the limited computational resources, we limit our focus to the validation split of TriviaQA Wikipedia questions. We construct a True-False dataset from two additional QA datasets, SQuAD 2.0 (Rajpurkar et al., 2018) and TruthfulQA (Lin et al., 2022), employing the same approach, which we describe

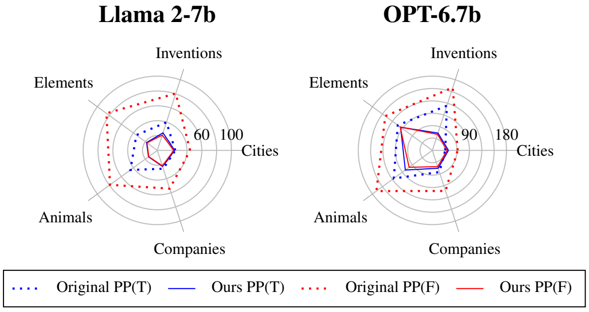

Figure 1: Comparison of average perplexity scores for Llama 2-7b and OPT-6.7b for the Original dataset by Azaria and Mitchell (2023) and our refined version. Lower perplexity indicates that the sentences are more likely to have been generated by the model. PP(T/F) denotes average perplexity of true/false sentences.

in Appendix E. For each LLM in our analysis, we set $K = 10$ and $\tau = 0.1$, and we discard the answers composed of less than 5 tokens to improve self-consistency. For annotation, we use **GPT-4o mini** with a 3-shot learning strategy, where three examples are provided. More details are included in Appendix C.1, and Table 8 reports the dataset statistics.

5.2 Impact of Perplexity-Based Sampling on SAPLMA Accuracy

We assess the impact of our perplexity-based sampling strategy by training the SAPLMA probe classifier separately on the original True-False dataset from Azaria and Mitchell (2023) and our refined version. As shown in Table 5, classifiers tested on our refined dataset achieve the highest accuracy in deeper layers. For Llama 2-7b, Layer 16 shows the most consistent results, achieving the highest average accuracy across all tested configurations when trained on our dataset. Although accuracy fluctuates across different topics, overall performance remains largely consistent between the training configurations. Comparing the results with the original results in Table 1, it emerges that classifiers trained on the original dataset achieve higher accuracy when tested on the same dataset. This is particularly evident in the 'Animals' topic. However, it is important to consider that the perplexity values in our refined dataset are significantly lower than those in the original dataset for both true and false statements (see 1). Additionally, our refined dataset features lower and more closely aligned perplexity scores between true and false statements, with false statements often yielding even lower perplexity values than true ones. This finding implies the validity of Azaria and Mitchell hypothesis, which states that **factuality information is encoded in LLM hidden states**, although it might not be immediately evident from the model predicted probabilities.

RQ1: *Are probing classifiers trained as in Azaria and Mitchell (2023) capable of generalizing to True-False sentences with similar perplexity?*

Surprisingly, the results between the probes trained on our refined dataset and the probes trained on the original dataset are mostly comparable. This proves that **the probes can generalize even when the train-test datasets have different perplexity**.

5.3 Generalization of SAPLMA on LLM-generated Sentences

We extend our experiments with the SAPLMA classifier to a new setting, where both training and evaluation are conducted on sentences generated by an LLM and sourced from TriviaQA, leveraging the procedure described in Section 4.2. Table 6 reports the performance of SAPLMA on a set of factual statements extracted from TriviaQA. The results indicate that the currently used probes are inadequate for factuality self-assessment in real-world scenarios, as the observed accuracy does not reach a noteworthy threshold. Furthermore, following the suggestion of Azaria and Mitchell, we optimize the classification threshold; however, this adjustment yields no significant improvement in performance. For completeness, the interested reader may find the result obtained on the same dataset when training on the original True-False dataset of Azaria and Mitchell in Appendix D. In addition, we report results on SQuAD 2.0 and TruthfulQA in Appendix E, which align closely with the results observed on TriviaQA, further supporting the consistency of our findings.

RQ2: *Can the same probes generalize to facts generated by LLMs?*

In summary, this experiment partially contradicts the findings of Azaria and Mitchell (2023): **the trained probes are not capable of providing good generalization to an LLM-generated dataset**, even when the accuracy threshold is tuned. The motivation could stem from the nature of the dataset, as TriviaQA contains open-domain questions that result in more nuanced facts than the ones in the original True-False dataset (Section 2) or the OPT-generated dataset (Section 3.2).

We believe that further research is needed to enhance the effectiveness of factuality self-assessment

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 5: Accuracy values obtained training SAPLMA on the original True-False dataset and on our refined version, then **tested on our refined version**. Results are shown for the Llama 2-7b and OPT-6.7b models. 'Orig.' denotes the 'original True-False dataset as training data, while 'Novel' denotes our version of the True-False dataset as training data. In **bold** we denote the best combination of layer/training dataset for each combination of model/topic.

Dataset		Threshold = 0.5					Optimal Threshold				
		last	28	24	20	16	last	28	24	20	16
billturnbull	Llama	.561	.576	.618	.621	.628	.547	.581	.618	.634	.639
	OPT	.547	.558	.591	.551	.537	.530	.558	.547	.530	.499
derby*	Llama	.568	.581	.575	.596	.617	.560	.565	.559	.571	.597
	OPT	.553	.564	.584	.587	.572	.562	.566	.583	.586	.580
quiz4free	Llama	.564	.547	.523	.561	.589	.548	.523	.521	.547	.573
	OPT	.559	.559	.575	.581	.560	.544	.530	.533	.546	.541
quizguy	Llama	.578	.585	.588	.607	.635	.576	.587	.601	.595	.637
	OPT	.579	.583	.589	.590	.584	.559	.555	.565	.559	.548
triviabug	Llama	.494	.518	.521	.525	.538	.508	.500	.530	.542	.545
	OPT	.620	.624	.607	.596	.528	.632	.635	.605	.598	.553
businessballs	Llama	.566	.558	.565	.574	.582	.558	.551	.555	.564	.575
	OPT	.559	.558	.578	.570	.553	.545	.547	.573	.562	.551
jetpunk	Llama	.587	.627	.620	.643	.654	.543	.580	.550	.554	.601
	OPT	.606	.612	.614	.596	.621	.618	.618	.621	.605	.619
odquiz	Llama	.551	.536	.546	.562	.573	.537	.525	.532	.550	.560
	OPT	.560	.573	.583	.583	.542	.551	.566	.571	.569	.521
quiz-zone	Llama	.556	.557	.558	.565	.611	.541	.549	.555	.569	.615
	OPT	.569	.570	.582	.592	.552	.537	.550	.535	.580	.508
quizballs	Llama	.603	.575	.572	.578	.571	.602	.561	.561	.570	.553
	OPT	.558	.565	.574	.571	.540	.550	.549	.564	.564	.539
quizwise	Llama	.560	.565	.579	.609	.618	.563	.551	.574	.605	.619
	OPT	.560	.563	.565	.577	.540	.555	.556	.552	.575	.537
sfquiz	Llama	.568	.554	.554	.559	.575	.564	.554	.559	.566	.583
	OPT	.584	.591	.589	.590	.547	.570	.582	.577	.573	.536
triviacountry	Llama	.536	.554	.559	.556	.566	.506	.530	.536	.524	.549
	OPT	.536	.551	.550	.587	.534	.449	.476	.472	.511	.472
wrexham**	Llama	.570	.563	.569	.553	.565	.567	.548	.561	.545	.566
	OPT	.548	.573	.578	.584	.554	.535	.552	.574	.579	.571
Average	Llama	.562	.564	.568	.579	.594	.552	.550	.558	.567	.587
	OPT	.567	.575	.583	.583	.555	.553	.560	.562	.567	.587

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

Table 6: Performance of SAPLMA on a fact dataset generated from TriviaQA. The original topic-wise leave-one-out strategy is adopted. Results are shown for the Llama 2-7b and OPT-6.7b models.

techniques, particularly in settings involving LLM-generated content. Promising research directions may be leveraging datasets that are closely aligned with the distribution of LLM-generated text and exploring alternative techniques such as uncertainty-aware classification.

6 Related Work

Probing techniques have become central for the layer-wise interpretation of deep learning models (Alain, 2016). This approach was then extended to Large Language Models (LLMs) to assess LLMs ability to encode syntax and semantics (Conneau et al., 2018; Tenney et al., 2019). Specifically, among the different applications, a technique that emerges is **self-evaluation**, defined as *a model's ability to assess the accuracy of its own outputs* (Kadavath et al., 2022). Among the various works, notable research by Kadavath et al. (2022) explored estimating a well-calibrated *P(True)* directly from output probabilities to reflect answer accuracy. Azaria and Mitchell (2023) demonstrate that LLMs are capable of detecting false claims in synthetic true-false datasets. Building upon this initial study, several works (Marks and Tegmark, 2023; Bürger et al., 2024; Levinstein and Herrmann, 2024) investigate the generalization capabilities of probes in detecting hallucinations in LLMs. While these studies offer valuable insights into how factual knowledge is encoded in LLMs, they stop short of examining whether LLMs can assess the factuality of their own generations. Instead, they primarily evaluate models on artificially constructed or perturbed datasets, rather than on claims produced by the LLMs themselves. Orgad et al. (2024) explore error detection and hallucination mitigation in Question Answering by probing the internal representations of question-answer pairs. Their study shows that while LLMs encode factuality signals, these signals do not generalize well across task-specific datasets, suggesting that factuality encoding is task-dependent rather than universal. Gekhman et al. (2025) show that LLMs can internally encode knowledge, detectable by probing the hidden states, yet still fail to express it in their generated output. Chen et al. (2024) introduce the INSIDE framework, which detects hallucinations using the EigenScore metric to measure self-consistency across multiple LLM outputs for a single input. Similarly, Zhang et al. (2024) leverage a probe model, PINOSE, trained via offline consistency-checking, to perform online hallucination detection. Although INSIDE and PINOSE primarily focus on self-consistency in internal representations across multiple generated responses, they do not interpret hidden states or look for explicit factuality encoding.

7 Conclusion and Future Work

In this paper, we investigate the factuality-encoding capabilities of LLMs. Our work replicates the methodology of Azaria and Mitchell (2023) to ensure reproducibility and extend their approach with two novel dataset construction strategies: perplexity-based negative sampling and fact generation based on QA datasets. We applied these strategies to analyze two open-source LLMs, and found that although the findings from previous studies are partially validated even on more challenging synthetic datasets, transferring these findings to LLM-generated datasets proves difficult. This study paves the way for more reliable LLM evaluations and offers practical guidelines for improving model transparency and trustworthiness in real-world applications.

Limitations

There are a few aspects of our study that could be explored further. We made every effort to check the datasets for inconsistencies, but a more thorough manual verification by human annotators would be beneficial for ensuring their robustness and minimizing potential biases. Additionally, our analysis is based on a limited set of models. While these models provide valuable insights, it is possible that larger or more complex models could demonstrate enhanced performance, particularly in the context of self-evaluation. Future work could expand on this, incorporating a wider range of models to investigate whether scalability can improve results.

Lastly, we highlight important considerations regarding the use of perplexity as a proxy for plausibility in relation to model knowledge. Recent findings suggest that LLMs often assign similar perplexity scores to both seen and unseen sentences (Duan et al., 2024), which may limit the reliability of perplexity in distinguishing between plausible and implausible content. This challenge intersects with broader issues of memorization and data provenance. While our work focuses on extending and refining prior approaches to factuality self-evaluation, we recognize that a rigorous assessment of perplexity effectiveness in this context is a critical avenue for future investigation. We therefore encourage further research into more robust and theoretically grounded measures of plausibility for LLM-generated content.

Ethical Statement

A major concern when working with LLMs is their tendency to generate factually inaccurate information. When training probe classifiers to assess factual accuracy, biases and beliefs from the LLM may transfer to the probe, potentially reinforcing cultural, demographic, or ideological biases in factuality self-evaluation. With careful design, probing techniques can be adapted not only to minimize bias but also to actively mitigate its consequences.

Acknowledgments

This study was supported by the MOST – Sustainable Mobility National Research Center funded by the European Union Next- GenerationEU (Italian National Recovery and Resilience Plan (NRRP) – M4C2, Investment 1.4 – D.D. 1033 17/06/2022, CN00000023 - CUP: D93C22000410001). We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). Patti Territoriali WP1, OVS: Fashion Retail Reloaded, Lutech Digitale 4.0, Natuzzi S.p.A. del Contratto di Sviluppo Industriale ai sensi dell'art. 9 del Decreto del Ministro dello Sviluppo Economico del 09.12.2014. EPANSA (FAIR) - Enhancing Personal Assistants with Neuro-Symbolic AI and Knowledge Graphs, funded by the European Union Next- GenerationEU (NRRP – M4C2, Investment 1.3, D.R. No. 123 of 16/01/2024, PE00000013, CUP: H97G22000210007). XAI4AMELIA - Implementation of Explainable Artificial Intelligence Technologies for Advancing Interoperability and Analysis in AMELIA, funded by the European Union Next- GenerationEU (NRRP – M4C2, Investment 1.3, D.D. 341 del 15/03/2022, CUP: J33C22002910001). Furthermore, this work has been carried out while Giovanni Servedio was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Politecnico di Bari.

References

Guillaume Alain. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Vito Walter Anelli, Giovanni Maria Biancofiore, Alessandro De Bellis, Tommaso Di Noia, and Eugenio Di Sciascio. 2022. Interpretability of BERT latent space through knowledge graphs. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 3806–3810. ACM.

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

Giovanni Maria Biancofiore, Dario Di Palma, Claudio Pomo, Fedelucio Narducci, and Tommaso Di Noia. 2025. Conversational user interfaces and agents. In *Human-Centered AI: An Illustrated Scientific Quest*, pages 399–438. Springer.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in llms. In *NeurIPS*.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *CoRR*, abs/2501.10970.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *CoRR*, abs/2401.01301.

Alessandro De Bellis, Vito Walter Anelli, Tommaso Di Noia, and Eugenio Di Sciascio. 2024. PRONTO: prompt-based detection of semantic containment patterns in MLMs. In *ISWC (2)*, volume 15232 of *Lecture Notes in Computer Science*, pages 227–246. Springer.

Dario Di Palma. 2023. Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *RecSys*, pages 1369–1373. ACM.

Dario Di Palma, Felice Antonio Merra, Maurizio Sfilio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. Do llms memorize recommendation datasets? a preliminary study on movielens-1m. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy July 13-18, 2025*. ACM.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? *CoRR*, abs/2402.07841.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Zorik Gekhman, Eyal Ben-David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. Inside-out: Hidden factual knowledge in llms. *CoRR*, abs/2503.15299.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL (1)*, pages 1601–1611. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario

Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Benjamin A Levinstein and Daniel A Herrmann. 2024. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pages 1–27.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL (1)*, pages 3214–3252. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *CoRR*, abs/2310.06824.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of LLM hallucinations. *CoRR*, abs/2410.02707.

Duy Khoa Pham and Bao Quoc Vo. 2024. Towards reliable medical question answering: Techniques and challenges in mitigating hallucinations in language models. *CoRR*, abs/2408.13808.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL (2)*, pages 784–789. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, and Yasmine Babaei et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Shriyash Upadhyay, Etan Ginsberg, and Chris Callison-Burch. 2023. Improving mathematics tutoring with A code scratchpad. In *BEA@ACL*, pages 20–28. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xiaokang Zhang, Zijun Yao, Jing Zhang, Kaifeng Yun, Jifan Yu, Juanzi Li, and Jie Tang. 2024. Transferable and efficient non-factual content detection via probe training with offline consistency checking. *arXiv preprint arXiv:2404.06742*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Appendix

A Dataset Statistics

A.1 True-False Dataset (Azaria and Mitchell, 2023) Statistics

Table 7 provides a summary of the number of sentences and the distribution of true and false statements across each topic in the original True-False dataset (Azaria and Mitchell, 2023) on which we base our replication experiments.

Dataset	Sentences	Source	(%) True
Cities	1458	SimpleMaps dataset	50
Inventions	876	Wikipedia's list of inventors	53
Chemical Elements	930	PubChem's periodic table	50
Animals	1008	National Geographic Kids	50
Companies	1200	Forbes Global 2000 List 2022: The Top 200	50
Scientific Facts	612	ChatGPT and human annotators	50

Table 7: True-False dataset categories, original sources, and label splits from (Azaria and Mitchell, 2023).

A.2 LLM-Generated Trivia Facts Dataset

Table 8 illustrates the number of annotated facts extracted from TriviaQA (Joshi et al., 2017) employing the procedure described in section 4.2. Statistics are presented for each of the 14 question sources. Table 9 reports examples of generated facts for the Llama 2-7b and OPT-6.7b models.

Dataset	Llama 2-7b		OPT-6.7b	
	Sentences	(%) True	Sentences	(%) True
triviacountry	118	49.15%	118	49.15%
wwwodquiz	700	49.71%	525	49.71%
triviabug	97	50.52%	88	51.14%
derby*	342	50.00%	325	50.00%
quiz-zone	187	48.13%	239	50.63%
businessballs	433	50.84%	443	49.43%
wrexham**	216	50.00%	276	50.72%
sfquiz	1054	49.80%	1085	50.39%
quizwise	565	49.91%	704	50.42%
billturnbull	216	49.54%	105	50.48%
jetpunk	139	51.80%	364	50.27%
quizballs	500	49.60%	488	50.82%
quizguy	240	51.67%	341	50.74%
quiz4free	171	49.12%	141	48.94%

Table 8: Summary of the dataset obtained by extracting factoid sentences from TriviaQA (Joshi et al., 2017) Wikipedia validation split, following the procedure described in Section 4.2.

B Perplexity-Based Refinement of the True-False Dataset (Azaria and Mitchell, 2023)

We base our dataset generation strategy described in Section 4 on the same tabular data employed by Azaria and Mitchell, which was made available

Model	Label	Sentence
Llama 2-7b	1	Arthur was married to Guinevere.
	1	Arthur's most famous wife was Guinevere.
	1	Guinevere was married to Arthur.
	1	Guinevere's husband was King Arthur.
	0	Lancelot was married to Guinevere.
	0	Sir Lancelot was married to Queen Guinevere.
	0	Sir Leonne was married to Queen Guinevere.
OPT-6.7b	0	Anakin Skywalker is Darth Vader's son.
	0	Darth Vader's son is Darth Vader.
	1	Darth Vader's son is Luke Skywalker, a member of the Rebel Alliance.
	1	Darth Vader's son is Luke Skywalker.
	1	Darth Vader's son is Luke.
	0	Darth Vader's son is known as Darth Vader.
	1	Luke's father is Darth Vader.

Table 9: Examples of true and false facts generated by Llama 2-7b and OPT-6.7b based on the questions in TriviaQA (Joshi et al., 2017), obtained following the procedure detailed in Section 4.2.

upon request. The properties used for sampling are analogous to the ones employed by the original authors in their dataset generation. However, for the *Cities* topic, we restrict the analysis to facts like "*<city> is city in <country>*." Differently from the original authors, we avoid generating facts such as "*<city> is the name of a city/country*" having only two possible values, potentially resulting in too many easy-to-classify samples. Table 10 reports examples of false sentences generated following the perplexity-based sampling procedure described in Section 4.1 and the Llama 2-7b model.

True Sentence	Original Negative	Generated Negative (Llama 2-7b)
The crocodile has a habitat of freshwater.	The crocodile has a habitat of **various.**	The crocodile has a habitat of **grassland/savanna.**
The zebra has distinctive black and white stripes, which may help deter flies and provide camouflage.	The zebra is a **fast swimmer** and can maintain high speeds for extended periods of time.	The zebra is the **fastest land animal**, reaching speeds up to 60-70 mph.
Tantalum has the symbol Ta.	Tantalum has the symbol **Cs.**	Tantalum has the symbol **Tm.**

Table 10: Examples of false sentences generated with the Llama 2-7b model, using the perplexity-based generation strategy described in Section 4.1.

C LLM-Generated Dataset Extraction Details

C.1 Factoid Answer Generation

Below is an example prompt used in the text generation pipeline for TriviaQA (Joshi et al., 2017). The 10 examples are sampled from the train Wikipedia split. The answers for the examples are manually crafted by looking at the available ground truth.

```
Question: Where in England was Dame Judi
     Dench born?
Answer: The English actress Dame Judi
    Dench was born in York, England.

Question: From which country did Angola
    achieve independence in 1975?
```

```
Answer: Angola achieved independence
    from Portugal in 1975.

Question: Which city does David Soul
    come from?
Answer: David Soul hails from Chicago,
    Illinois.

Question: Who won Super Bowl XX?
Answer: The Chicago Bears won Super Bowl
     XX.

Question: Which was the first European
    country to abolish capital
    punishment?
Answer: Norway was the first European
    country to abolish capital
    punishment.

Question: In which country did the
    widespread use of ISDN begin in
    1988?
Answer: The widespread use of ISDN began
     in Japan in 1988.

Question: What is Bruce Willis' real
    first name?
Answer: Bruce Willis' real first name is
     Walter.

Question: Which William wrote the novel
    Lord of the Flies?
Answer: The William who wrote Lord of
    the Flies was William Golding.

Question: How is Joan Molinsky better
    known?
Answer: Joan Molinsky is better known as
     Joan Rivers.

Question: In which branch of the arts is
     Patricia Neary famous?
Answer: Patricia Neary is famous in the
    field of ballet.
```

To generate responses, the model is provided with a continuation prompt such as:

```
Question: Ford Prefect came from a star
    in which constellation?
Answer:
```

The model then generates a response, adhering to the pattern set by the previous examples. Responses are filtered based on length and uniqueness to ensure meaningful and high-quality outputs.

The generation process uses the following hyperparameters for all the employed models (i.e. Llama 2-7b and OPT-6.7b):

- `max_new_tokens`: 128
- `top_k`: 50
- `top_p`: 0.95
- `temperature`: 1.0
- `num_return_sequences`: 10

C.2 Factoid Answer Annotation

Below is an example prompt to **GPT-4o mini** (2024-08-01), the LLM that we use as the oracle for the procedure described in Section 4.2. In this setup, we employ a 3-shot learning strategy, where the model is provided with three examples to establish a clear pattern of judgment. These examples demonstrate the task of evaluating answers against a given ground truth. The model is then tasked with using this context to evaluate a new answer to a question, based on the same criteria. In our experiments, we employ a temperature setting of 1.

```
SYSTEM:

You are a judge and your role is to
    judge whether the provided answer is
     correct for the given question,
    based on the provided ground truth.
    Answer with a 1 if the answer is
    correct and 0 if the answer is
    incorrect.
Here are a few examples:

USER:

Question: Who was the next British Prime
     Minister after Arthur Balfour?
Answer: Arthur Balfour was followed by
    David Lloyd George.
Ground truth: "{'aliases': ['Sir Henry
    Campbell-Bannerman', 'Campbell-
    Bannerman', 'Campbell Bannerman', '
    Sir Henry Campbell Bannerman', '
    Henry Campbell Bannerman', 'Henry
    Campbell-Bannerman'], '
    normalized_aliases': ['henry
    campbell bannerman', 'sir henry
    campbell bannerman', 'campbell
    bannerman'], '
    matched_wiki_entity_name': '', '
    normalized_matched_wiki_entity_name
    ': '', 'normalized_value': 'campbell
     bannerman', 'type': '
    WikipediaEntity', 'value': 'Campbell
    -Bannerman'}"
Evaluation:

ASSISTANT:

0

USER:

Question: Who had a 70s No 1 hit with
    Kiss You All Over?
Answer: The band Exile had a 70s No 1
    hit with Kiss You All Over.
Ground truth: "{'aliases': ['Internal
    exile', 'Exiles', 'Transported for
    life', 'Exile (politics and
    government)', 'Voluntary exile', '
    Sent into exile', 'Exile and
    Banishment', 'Self-exile', 'Forced
```

```
    exile', 'Exile', 'Exile in Greek
    tragedy', 'Banish', 'Banishment'], '
    normalized_aliases': ['exiles', '
    voluntary exile', 'forced exile', '
    banish', 'self exile', 'exile
    politics and government', 'exile in
    greek tragedy', 'sent into exile', '
    banishment', 'transported for life',
     'exile', 'internal exile', 'exile
    and banishment'], '
    matched_wiki_entity_name': '', '
    normalized_matched_wiki_entity_name
    ': '', 'normalized_value': 'exile',
    'type': 'WikipediaEntity', 'value':
    'Exile'}"
Evaluation:

ASSISTANT:

1

USER:

Question: Which common mineral is used
    to make casts, moulds, blackboard
    chalk and plaster of Paris?
Answer: The common mineral used to make
    casts, moulds, blackboard chalk and
    plaster of Paris is calcium
    carbonate.
Ground truth: "{'aliases': ['CaSO4.2H2O
    ', 'Gypsum', 'Calcium sulfate
    dihydrate', 'CaSO4*2H2O', 'Gipsum'],
     'normalized_aliases': ['calcium
    sulfate dihydrate', 'caso4 2h2o', '
    gipsum', 'caso4.2h2o', 'gypsum'], '
    matched_wiki_entity_name': '', '
    normalized_matched_wiki_entity_name
    ': '', 'normalized_value': 'gypsum',
     'type': 'WikipediaEntity', 'value':
     'Gypsum'}"
Evaluation:

ASSISTANT:

0
```

D Additional Experiments on the LLM-Generated Dataset

Table 11 reports the performances of the SAPLMA classifier trained on the original True-False dataset by Azaria and Mitchell (2023) and tested on the dataset generated from TriviaQA. Similarly to Table 6, SAPLMA does not generalize well over LLM-generated facts. Moreover, tuning an optimal threshold did not provide solid enhancements.

E Experiments on Additional LLM-Generated Datasets

We extended our LLM-generated dataset construction approach (Section 4.2) to TruthfulQA (Lin et al., 2022) and SQuAD 2.0 (Rajpurkar et al., 2018) to further support our findings.

Dataset		Threshold = 0.5					Optimal Threshold				
		last	28	24	20	16	last	28	24	20	16
billturnbull	Llama	.579	.560	.560	.593	.648	.605	.553	.632	.632	.691
	OPT	.543	.533	.533	.486	.476	.527	.500	.486	.500	.500
derby*	Llama	.556	.576	.550	.544	.602	.542	.529	.529	.554	.575
	OPT	.526	.548	.554	.563	.535	.535	.583	.570	.579	.561
quiz4free	Llama	.602	.573	.544	.538	.608	.575	.600	.525	.550	.642
	OPT	.525	.489	.511	.525	.504	.545	.535	.556	.576	.495
quizguy	Llama	.608	.571	.546	.567	.571	.601	.595	.583	.577	.565
	OPT	.557	.557	.587	.569	.557	.586	.594	.552	.548	.552
triviabug	Llama	.412	.619	.526	.598	.577	.485	.471	.471	.544	.544
	OPT	.557	.602	.671	.614	.500	.629	.597	.629	.597	.532
businessballs	Llama	.580	.577	.559	.575	.589	.586	.605	.546	.563	.592
	OPT	.564	.578	.555	.587	.521	.585	.579	.585	.547	.537
jetpunk	Llama	.612	.590	.583	.619	.640	.582	.561	.520	.592	.673
	OPT	.569	.571	.593	.604	.566	.620	.631	.631	.631	.635
odquiz	Llama	.546	.536	.537	.559	.564	.527	.522	.547	.563	.571
	OPT	.511	.543	.591	.579	.552	.565	.535	.592	.571	.565
quiz-zone	Llama	.578	.519	.588	.562	.578	.611	.534	.534	.534	.557
	OPT	.544	.577	.603	.611	.586	.542	.518	.554	.601	.613
quizballs	Llama	.610	.568	.592	.582	.582	.617	.563	.583	.557	.586
	OPT	.535	.578	.559	.594	.549	.532	.564	.512	.599	.576
quizwise	Llama	.572	.588	.570	.572	.586	.581	.598	.611	.616	.616
	OPT	.540	.550	.574	.568	.551	.580	.550	.554	.582	.576
sfquiz	Llama	.530	.533	.528	.538	.560	.545	.581	.584	.570	.575
	OPT	.531	.546	.568	.590	.545	.595	.597	.599	.611	.553
triviacountry	Llama	.602	.636	.602	.602	.619	.639	.602	.639	.590	.578
	OPT	.525	.542	.585	.551	.525	.494	.602	.506	.578	.446
wrexham**	Llama	.532	.519	.528	.574	.583	.474	.526	.533	.566	.579
	OPT	.500	.533	.594	.583	.525	.572	.552	.562	.608	.567
Average	Llama	.566	.569	.558	.573	.593	.569	.560	.560	.572	.596
	OPT	.538	.553	.577	.573	.535	.565	.567	.563	.581	.551

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

Table 11: Accuracy values obtained training SAPLMA on the original True-False dataset and testing on our facts dataset generated from TriviaQA. The original topic-wise leave-one-out strategy is adopted. Results are shown for the Llama 2-7b and OPT-6.7b models.

E.1 LLM-Generated Facts from TruthfulQA

Model	Total Samples	True Samples (%)
Llama 2-7b	842	49.76%
OPT-6.7b	331	49.85%

Table 12: Summary of the dataset obtained by extracting factoid sentences from TruthfulQA (Lin et al., 2022), following the procedure described in Section 4.2.

Following the same procedure detailed in Section 4.2, we extract true and false facts from TruthfulQA. Table 12 reports the statistics of the LLM-generated datasets constructed with Llama 2-7b and OPT-6.7b. Table 13 presents the performance of the SAPLMA classifier when both trained and tested on the TruthfulQA dataset. Additionally, Table 14 shows the results obtained by SAPLMA on TruthfulQA after being trained on the original True-False dataset. Since the dataset lacks a topic-based split, we adopt a 50-50 holdout strategy for evaluation. The results on the TruthfulQA dataset

	OPT-6.7b		Llama 2-7b	
	Random Split 1	Random Split 2	Random Split 1	Random Split 2
last	0.487	0.488	0.491	0.504
28	0.477	0.495	0.498	0.506
24	0.478	0.513	0.507	0.507
20	0.512	0.535	0.518	0.509
16	0.493	0.492	0.520	0.514

Table 13: Performance of SAPLMA on a fact dataset generated from TruthfulQA. A 50-50 holdout strategy is adopted. Results are shown for the Llama 2-7b and OPT-6.7b models.

	OPT-6.7b		Llama 2-7b	
	Random Split 1	Random Split 2	Random Split 1	Random Split 2
last	0.473	0.560	0.518	0.539
28	0.515	0.566	0.520	0.514
24	0.503	0.476	0.518	0.501
20	0.455	0.542	0.499	0.525
16	0.479	0.500	0.485	0.530

Table 14: Accuracy values obtained training SAPLMA on the original True-False dataset and testing on our facts dataset generated from TruthfulQA. A 50-50 holdout strategy is adopted. Results are shown for the Llama 2-7b and OPT-6.7b models.

support our findings: the probe classifier fails to generalize effectively to LLM-generated datasets.

E.2 LLM-Generated Facts from SQuAD 2.0

For the SQuAD 2.0 dataset, we limit our dataset construction procedure (Section 4.2) to the validation split and the five most common topics, as the large number of questions makes applying our procedure to the full dataset computationally impractical due to time constraints. Table 16 presents the performance of the SAPLMA classifier when trained and tested on the SQuAD 2.0 dataset. Table 17, instead, reports the results obtained on SQuAD 2.0 after training on the original True-False dataset from Azaria and Mitchell (2023).

Topic	Model	Sentences	(%) True
Economic_inequality	Llama 2-7b	224	50.45%
	OPT-6.7b	135	50.37%
Immune_system	Llama 2-7b	167	49.70%
	OPT-6.7b	72	50.00%
Rhine	Llama 2-7b	120	50.83%
	OPT-6.7b	59	47.46%
Warsaw	Llama 2-7b	140	51.43%
	OPT-6.7b	102	50.00%
Yuan Dynasty	Llama 2-7b	109	49.54%
	OPT-6.7b	53	52.83%

Table 15: Summary of the dataset obtained by extracting factoid sentences from SQuAD 2.0 (Rajpurkar et al., 2018), following the procedure described in Section 4.2.

Dataset		last	28	24	20	16
Economic_inequality	Llama	0.525	0.510	0.515	0.492	0.502
	OPT	0.517	0.516	0.499	0.496	0.464
Immune_system	Llama	0.486	0.490	0.503	0.502	0.547
	OPT	0.436	0.439	0.401	0.438	0.435
Rhine	Llama	0.525	0.520	0.515	0.524	0.526
	OPT	0.484	0.485	0.522	0.547	0.521
Warsaw	Llama	0.558	0.542	0.545	0.546	0.555
	OPT	0.521	0.531	0.545	0.544	0.509
Yuan_dynasty	Llama	0.509	0.522	0.514	0.529	0.544
	OPT	0.544	0.560	0.512	0.542	0.512

Table 16: Performance of SAPLMA on a fact dataset generated from SQuAD 2.0, restricted to the validation set and the top-5 most popular topics. The original topic-wise leave-one-out strategy is adopted. Results are shown for the Llama 2-7b and OPT-6.7b models.

		last	28	24	20	16
Economic_inequality	Llama	0.558	0.531	0.589	0.531	0.545
	OPT	0.600	0.600	0.548	0.533	0.533
Immune_system	Llama	0.569	0.593	0.593	0.575	0.599
	OPT	0.611	0.611	0.611	0.708	0.542
Rhine	Llama	0.467	0.492	0.475	0.442	0.525
	OPT	0.525	0.559	0.593	0.593	0.576
Warsaw	Llama	0.614	0.579	0.607	0.600	0.593
	OPT	0.48	0.500	0.539	0.598	0.549
Yuan_dynasty	Llama	0.523	0.523	0.541	0.505	0.550
	OPT	0.453	0.547	0.491	0.528	0.547

Table 17: Accuracy values obtained training SAPLMA on the original True-False dataset and testing on the facts dataset generated from SQuAD 2.0. Results are shown for the Llama 2-7b and OPT-6.7b models.

The results on the SQuAD 2.0 dataset further support our conclusion: existing probes fall short in accurately assessing factuality in real-world settings. Across all cases, accuracy remains below a meaningful threshold, highlighting the limited reliability of these methods for self-assessment in scenarios involving LLM-generated sentences.

F Examples of SAPLMA Predictions

The tables in this appendix provide illustrative examples that complement the quantitative results discussed in the main text. Specifically, Table 18 presents sentence-level predictions on facts generated from the TriviaQA dataset using Llama 2-7b. In this case, the probe classifier is trained adopting a topic-wise leave-one-out strategy, as in Table 6. Table 19 shows additional predictions obtained using the probe classifier, which is trained on all but one topic and tested on our refined version of the held-out topic (as in the *Orig.* setting of Table 5).

Sentence	Predicted	GT	Confidence
Sir John Suckling was the first poet to be buried at Poet's Corner in London's Westminster Abbey.	False	False	0.373
The first poet to be buried at Poet's Corner in London's Westminster Abbey was Geoffrey Chaucer.	True	True	0.551
The poet who was the first to be buried in Westminster Abbey was Geoffrey Chaucer.	False	True	0.456
Roquefort cheese is made from sheep's milk.	True	True	0.794
Roquefort cheese is made from cow's milk.	True	False	0.851
The milk of sheep is used to make 'Roquefort Cheese'.	False	True	0.488
South Korea held its first 'Grand Prix' motor race in 1999.	False	False	0.291
South Korea held its first 'Grand Prix' motor race in 2010.	False	True	0.305
South Korea hosted its first 'Grand Prix' motor race in 1966.	False	False	0.182

Table 18: Examples of sentence classification on facts generated from the TriviaQA dataset using Llama 2-7b. The original topic-wise leave-one-out strategy is adopted. Correct predictions are highlighted in green, while incorrect ones are shown in red.

Sentence	Predicted	GT	Confidence
Tin has the symbol Sn.	True	True	0.990
Mercury is a liquid at room temperature and used in thermometers and some electrical switches.	False	True	0.113
Fluorine is in the Alkaline earth metal group.	False	False	0.012
Tellurium has the atomic number of 82.	True	False	0.998

Table 19: Examples of predictions using SAPLMA and Llama 2-7b. In these examples, the probe classifier is trained on the original version of the True-False dataset (Azaria and Mitchell, 2023) (excluding the *element* topic) and tested on our refined version of the *element* topic. Correct predictions are highlighted in green, while incorrect ones are shown in red.

The code is provided in our `GitHub Repository`.