Characterizing Truthfulness in Large Language Model Generations with Local Intrinsic Dimension

Fan Yin [1] **Jayanth Srinivasa** [2] **Kai-Wei Chang** [1]

Abstract

We study how to characterize and predict the truthfulness of texts generated from large language models (LLMs), which serves as a crucial step in building trust between humans and LLMs. Although several approaches based on entropy or verbalized uncertainty have been proposed to calibrate model predictions, these methods are often intractable, sensitive to hyperparameters, and less reliable when applied in generative tasks with LLMs. In this paper, we suggest investigating internal activations and quantifying LLM's truthfulness using the local intrinsic dimension (LID) of model activations. Through experiments on four question answering (QA) datasets, we demonstrate the effectiveness of our proposed method. Additionally, we study intrinsic dimensions in LLMs and their relations with model layers, autoregressive language modeling, and the training of LLMs, revealing that intrinsic dimensions can be a powerful approach to understanding LLMs. Code is available at: https://github.com/fanyin3639/LID-HallucinationDetection.

1. Introduction

Large language models (LLMs) have demonstrated remarkable effectiveness in various generative natural language processing (NLP) tasks, including QA, summarization, and dialogue (Touvron et al., 2023a; Chowdhery et al., 2023; OpenAI, 2023). However, deploying LLMs to more high-stakes scenarios remains limited due to their tendency to provide plausible but untruthful answers, even when they are uncertain, also known as hallucinations (Ji et al., 2023).

[1]Department of Computer Science, University of California, Los Angeles, LA, U.S.A. [2]Cisco Research, U.S.A. Correspondence to: Fan Yin <fanyin20@cs.ucla.edu>, Kai-Wei Chang <kwchang@cs.ucla.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

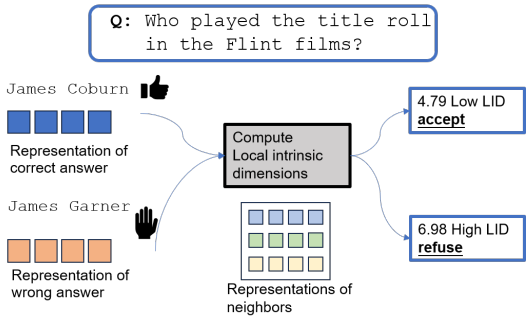

Figure 1. Detecting hallucinations with LIDs. LLM representations of correct answers have smaller intrinsic dimensions.

Hence, characterizing and eliciting the truthfulness of model outputs is a crucial step towards constructing more reliable LLMs and building user trust in models (Bommasani et al., 2021; Kadavath et al., 2022; Zou et al., 2023).

Despite its importance, little is known about which information within models most accurately characterizes their truthfulness. A mainstream of work approaches this through logit-level entropy-based uncertainty (Gal & Ghahramani, 2016; Malinin & Gales, 2020; Kuhn et al., 2022; Duan et al., 2023) or verbalized uncertainty (Kadavath et al., 2022; Zhou et al., 2023; Tian et al., 2023). However, computing uncertainty is limited to classification tasks and becomes intractable for generative tasks due to infinite output space. Moreover, extracting truthfulness only at the output layer inevitably loses substantial information, leading to sub-optimal performance.

Other approaches train linear probes to discover truthfulness directions in model internal representations (Azaria & Mitchell, 2023; Li et al., 2023; Burns et al., 2022; Zou et al., 2023; Marks & Tegmark, 2023). However, these truthful directions do not always exist and can vary significantly due to tasks, the layers being used, and the styles of prompts. Therefore, determining whether a direction would be beneficial for the task at hand can be cumbersome.

In this paper, we delve into the internal representations, which have been shown to preserve more information and geometric characteristics. Instead of seeking truthful directions for each task, we leverage a more principled and

generalizable feature to detect hallucinations: the discrepancy in local intrinsic dimension (LID) (Levina & Bickel, 2004; Gomtsyan et al., 2019) of model activations. LID reflects the minimal number of activations required to characterize the current point without significant information loss. A higher LID means that the current point lies in a more complicated manifold and vice verse. LLM representations, which are often high-dimensional vectors (e.g., 4,096 for Llama-2-7B, Touvron et al. 2023b) are commonly believed to lie in lower-dimensional manifolds because of the inductive bias of the model and the natural structure in human language (Marks & Tegmark, 2023). We hypothesize that truthful outputs, being closer to natural language, are more structured and have smaller LIDs. On the other hand, an untruthful continuation of a prompt hallucinated by the model itself would mix human (prompt) and complex model (continuation) distribution, leading to larger LIDs. The discrepancy in LID would thus serve as a strong signal to assess whether an output is truthful or not.

More specifically, our method is based on the well-established maximum likelihood estimation (MLE) method (Levina & Bickel, 2004) but proposes a simple yet effective correction to 1) accommodate the non-linearity in language representations (Gomtsyan et al., 2019); 2) select the optimal set of representations. MLE approximates the count of neighbors surrounding the current sample with a Poisson process parameterized by the LID. It inherently supports estimation for an individual sample while other estimators mostly consider intrinsic dimension as a property of the whole dataset. Our improvements enable more accurate estimations of LIDs in representations.

Experiments with the Llama-2 (Touvron et al., 2023b) family on four QA tasks prove the advantage of using LID methods over uncertainty methods, achieving an improvement of 8% under AUROC. Compared with representation-level methods like linear probes and t-SNE (Van der Maaten & Hinton, 2008), we show that our method is more powerful while other methods fail to discover truthful directions. Further ablation study shows that we could even leverage out-of-distribution samples as neighbors to estimate the LIDs, demonstrating the generalizability of our methods.

We further conduct a series of analyses on the intrinsic dimensions in LLMs, revealing several intriguing properties beyond its relations to hallucinations. Overall, we believe intrinsic dimension is an insightful and powerful feature for understanding LLMs. Below are our findings:

- Similar to the findings by Ansuini et al. (2019) on image data, we observe a 'hunchback' shape in the intrinsic dimension of language generations: the intrinsic dimension values increase in the first few layers and then gradually decrease. The hallucination detection performance curve follows a similar shape to the intrinsic dimension value curve but is 'shifted behind' by one or two layers;
- We verify our hypothesis that mixing human and model distributions increases intrinsic dimension by controlling where the 'mixing' happens. We show that the intrinsic dimensions for human answers are consistently lower than untruthful model outputs at every position, exhibiting a sharp decrease when the answer approaches the end;
- In addition to frozen foundation model, we are curious about how instruction tuning, a technique widely adopted to align LLMs, impacts intrinsic dimensions. We find that as the instruction tuning progresses, the intrinsic dimension of LLMs' representations tends to increase. Furthermore, intrinsic dimensions correlate with the generalization performance of the model.

2. Related Work

Characterizing Model Truthfulness As LLMs improve, it is increasingly crucial to ensure their safety and truthfulness (Hendrycks et al., 2021; Bommasani et al., 2021). An important technique is to detect incorrect model outputs so that users can decide when not to trust those outputs (Kamath et al., 2020; Ren et al., 2022). Existing techniques towards this goal mainly fall into three categories: 1) entropy-based uncertainty estimation (Malinin & Gales, 2020; Kuhn et al., 2022; Duan et al., 2023; Lin et al., 2023). However, those approximations are not accurate for LLMs since the output space of LLMs is too large; 2) Verbalized uncertainty (Kadavath et al., 2022; Tian et al., 2023; Zhou et al., 2023; Xiong et al., 2023), i.e., directly asking LLMs to judge their answers. It typically involves extra training as models are not pre-trained with this objective; 3) Probing truthfulness direction (Zou et al., 2023; Azaria & Mitchell, 2023; Li et al., 2023; Burns et al., 2022) that tries to find a truthfulness direction in model representations for a specific dataset. The obtained direction usually does not generalize well. We propose to use LIDs to detect hallucinations, which leverage geometric information that uncertainty-based methods ignore and are more generalizable than truthful directions.

(Local) Intrinsic Dimension in Neural Models Neural models are commonly believed to be redundant in terms of their parameters and representations (Birdal et al., 2021). Approaches to estimate intrinsic dimension of data manifolds (Levina & Bickel, 2004; Amsaleg et al., 2015; Facco et al., 2017b) have been applied to understand the structure in models. For example, Ma et al. (2018) use differences in LID values to characterize adversarial image data; Ansuini et al. (2019); Pope et al. (2020); Birdal et al. (2021) study the relation between intrinsic dimensions and generalization ability of models. Most related to ours, Tulchinskii et al. (2023) train classifiers using ID values to identify AI-generated texts in multiple languages. Different from previous work, we focus more on LID for each individual

sample, and use LID to identify incorrect model outputs.

3. LID for Characterizing Truthfulness

In this section, we formulate the problem of characterizing the truthfulness of model outputs. We review the MLE framework for LID and introduce the modifications to account for the LLMs' representations.

3.1. Problem Setup

Consider an L-layer causal LLM M that takes a sequence of N tokens $X = [x_1, x_2, \ldots, x_N]$ as input, and generates a sequence of O-token continuation as output $M(X) = [x_{N+1}, x_{N+2}, \ldots, x_{N+O}]$. $M(X)$ is generated in an autoregressive manner, where each $x_{N+i}, i \in [1, \ldots, O]$ is sampled from a distribution over the model vocabulary $\mathcal{V}$, conditioned on the prefix $[x_1, x_2, \ldots, x_N, x_{N+1}, \ldots, x_{N+i-1}]$:

$$\mathbf{X}_{Li} = (M_L \circ M_{L-1} \circ \cdots \circ M_0)([x_1, \ldots, x_{N+i-1}]),$$

$$p(x_{N+i} | [x_1 \ldots, x_{N+i-1}]) = \text{softmax}(W\mathbf{X}_{Li} + b),$$

where M_j, $j \in [1 \ldots L]$ is the j-th layer of the LLM M. M_0 is the embedding layer. W, b are the output projection weights and bias. For later sections, we use $\mathbf{X}_{ji} \in \mathcal{R}^D$ to denote the j-th layer representation for the i-th continuation token x_{N+i}, which is a D-dimensional vector. We denote the probability distribution over a single token x_{N+i} as $p(x_{N+i}|\cdot)$, and the whole sequential output $M(X)$ $p(M(X)|\cdot)$, respectively.

For a specific task with n points $\mathcal{D} = \{X^1, \ldots, X^n\}$, we aim to predict the truthfulness of each corresponding generation $\{M(X^1) \ldots M(X^n)\}$ before knowing the ground truth. Note that the truthfulness criteria might vary based on the task being considered, which might be string matching, semantic similarity, or any other human-based metrics. We use $\left\{\hat{Y}^1 \ldots \hat{Y}^n\right\}$ to denote the ground truth, and $s\left(M(X^i), \hat{Y}^i\right) \in \{0, 1\}$ to denote the indicator function for whether an input-output pair $(X^i, M(X^i))$ can be considered truthful. The goal of this paper is to propose a characterizing feature that accurately reflects $s\left(M(X^i), \hat{Y}^i\right)$. While previous works on uncertainty estimation mostly obtain the feature from the final predictive distribution $p(M(X)|\cdot)$, we instead propose to explore the LID of intermediate representations $\mathbf{X}_{ji}$.

3.2. MLE Estimator for LID

Here, we review the core idea of the MLE estimator for LIDs (Levina & Bickel, 2004). Notice that while there are other estimators available for intrinsic dimension (Costa et al., 2005; Facco et al., 2017a; Campadelli et al., 2015), MLE is specifically tailored for estimating 'local' intrinsic dimension of an individual point, making it well-suited for our application, unlike many others that estimate 'global' intrinsic dimension.

For the representation[1] of a data point in $\mathcal{D}$, $\mathbf{X}^i$, the MLE estimator, as its name suggests, conducts a maximum likelihood estimation of the count of its neighbors using a Poisson process, where the rate of the Poisson process is parameterized by the intrinsic dimension m. Formally, it considers the T nearest neighbors of $\mathbf{X}^i$ in $\mathcal{D}$, $\left\{\mathbf{X}^{i1}, \ldots, \mathbf{X}^{iT}\right\}$, and a ball of radius R centered at $\mathbf{X}^i$, $S_{\mathbf{X}^i}(R)$. The count of neighbors inside balls of varying radius $0 < t < R$ can be expressed by a binomial process as follows:

$$N(t, \mathbf{X}^i) = \sum_{k=1}^{T} \mathbb{I}\left\{\mathbf{X}^{ik} \in S_{\mathbf{X}^i}(t)\right\}.$$

Levina & Bickel (2004) propose to approximate the above with a Poisson process of a certain rate $\lambda(t)$. According to the definition of λ, if we assume the density f is approximately constant around $\mathbf{X}^i$, and the volume V expands proportionally to t^m, i.e., $V = V_m t^m$ where m is the intrinsic dimension, we will have

$$\lambda(t) = f\frac{dV}{dt} = fV_m\, m\, t^{m-1}.$$

Then, the log-likelihood of the Poisson process can be written as a function of the intrinsic dimension m and $\theta = \log f$.

$$L(m, \theta) = \int_0^R \log \lambda(t)\, dN_\lambda(t) - \int_0^R \lambda(t)\, dt. \quad (1)$$

Maximizing the log-likelihood in Eq. 1, we have the following formula for calculating the intrinsic dimension m:

$$m(R, \mathbf{X}^i) = \left(\frac{1}{N(R, \mathbf{X}^i)} \sum_{j=1}^{N(R, \mathbf{X}^i)} \log \frac{R}{Q_j}\right)^{-1}, \quad (2)$$

where Q_j, $j = 1, \ldots T$ is the Euclidean distance to the j-th nearest neighbor. The numerical calculation of Eq. 2 can be further simplified as:

$$m(\mathbf{X}^i) = \left(\frac{1}{T-1} \sum_{j=1}^{T-1} \log \frac{Q_T}{Q_j}\right)^{-1}. \quad (3)$$

3.3. Layer Selection and Distance-aware MLE

In the previous section, we discuss how to calculate LIDs for a representation $\mathbf{X}^i$. When it comes to LLMs, two challenges arise: 1) there will be a D-dimensional representation

[1] We use the representation from a middle layer and the last token, elaborated in Section3.3. For simplicity, we omit the subscripts for its position and layer.

for each position at each layer, making it hard to select the optimal representation to use; 2) MLE assumes constant density function f, which is unlikely to hold for causal LLMs on complicated real data. Next, we discuss our solutions for addressing those issues.

Layer Selection As mentioned earlier, LLMs generate a D-dimensional representation for each token at each layer. We select the token at the last position of $\mathbf{X}^i$, i.e., $\mathbf{X}^i_{-1}$ as that representation contains all pertinent information from preceding positions. This strategy aligns with other works on probing truthful directions like Zou et al. (2023).

However, when it comes to layer selection, our empirical evidence indicates that the representations from the last layer might not yield the most informative feature. As in Figure 3, the performance of predicting truthfulness with LIDs correlates well with the absolute value of summed LIDs over the test set, but exhibits a shift for one or two layers. Based on this observation, we propose to empirically selecting the layer l with the following criteria:

$$l = \text{argmax}_l \sum_{i=1}^{n} m\left(\mathbf{X}^i_{l\{-1\}}\right) + 1.$$

Distance-aware MLE To mitigate the non-uniformity of density when applying MLE of the Poisson process, a common practice involves adjusting the rate $\lambda(t)$ of the Poisson process. Gomtsyan et al. (2019) suggest to replace the original rate $\lambda(t) = fV_m m t^{m-1}$ as $\hat{\lambda}(t) = fV_m m t^{m-1} + t^m V_m \delta(t)$, where $\delta(R)$ is a correction function bounded by some geometric properties of the manifold of $\mathbf{X}^i$. For the exact form of $\delta(R)$, see Gomtsyan et al. (2019). We will review the steps after the correction.

With the new rate, maximizing the log-likelihood in Eq. 1 we will have:

$$\hat{m}\left(R, \mathbf{X}^i\right) = m\left(R, \mathbf{X}^i\right)\left(1 + \delta(R)\frac{R^2}{N\left(R, \mathbf{X}^i\right)}\right). \quad (4)$$

Then, with Taylor expansion on the second term in Eq. 4, we can calculate the correction with the following polynomial regression:

$$\hat{m}\left(R, \mathbf{X}^i\right) = m\left(R, \mathbf{X}^i\right) + \sum_{j=1}^{l} \zeta_j R^j + \Theta\left(R^{l+1}\right). \quad (5)$$

The new steps will be to estimate ζ_j, and use the zero-order term as the estimated $\hat{m}\left(\mathbf{X}^i\right)$:

$$m\left(\mathbf{X}^i\right) = \hat{m}\left(\mathbf{X}^i\right) + \sum_{j=1}^{l} \zeta_j Q_T^j + \Theta\left(Q_T^{l+1}\right). \quad (6)$$

Solving the polynomial regression in Eq. 6 can be conducted by minimizing the weighted least squared errors. We follow Gomtsyan et al. (2019) to use bootstrapping of $\mathcal{D}$: $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_p$. We calculate the average of LIDs and distances, as well as the variance of LIDs. In practice, we find that solely using the bootstrapping process brings improvements of the truthfulness detection performance.

$$\bar{Q}_T = \frac{1}{p}\sum_{i=1}^{p} Q_{Tp},\ \bar{m}\left(\mathbf{X}^i\right) = \frac{1}{p}\sum_{i=1}^{p} m\left(\mathbf{X}^i\right)_p,$$

$$\sigma(m(X^i)) = \frac{1}{p}\sum_{i=1}^{p}\left(m\left(\mathbf{X}^i\right)_p - \bar{m}\left(\mathbf{X}^i\right)\right)^2.$$

Finally, the heteroskedastic weighted polynomial regression is minimized to find the final LID over different numbers of neighbors T:

$$\min \sum_{T=T_1}^{T_2} \frac{1}{\sigma(m(\mathbf{X}^i))}\left(\hat{m}\left(\mathbf{X}^i\right) - m(\mathbf{X}^i) - \sum_{j=1}^{l}\zeta_j Q_T^j\right)^2.$$

We call the method with distance-aware MLE estimation as *LID-GeoMLE*, and the vanilla MLE method *LID-MLE*.

Clearly, for both LID-MLE and LID-GeoMLE, the estimated LID is a function of hyperparameters T, the number of neighbors, and n, the dataset size. While a small T and n provide an estimation of LIDs with large variance, a T or n that is too large will break the condition of local balls and neighbors. We elaborate in Section 4.3 the effects of them in the performance.

4. Experiments

In this section, we empirically demonstrate the effectiveness of using LID to predict the truthfulness of model outputs, outperforming uncertainty-based methods and classifiers trained to predict truthfulness.

Datasets & Models We consider four generative QA tasks: TriviaQA (Joshi et al., 2017), CoQA (Reddy et al., 2019), HotpotQA (Yang et al., 2018), and TydiQA-GP (English) (Clark et al., 2020). These datasets cover different formats of QA including open-book (CoQA), closed-book (TriviaQA, HotpotQA), and reading comprehension (TydiQA-GP) and several capacities of LLMs. For each of the datasets, we generate outputs for 2,000 samples from the validation sets and test the methods with those samples.

We evaluate with the decoder-only Transformer-based Llama-2 (Touvron et al., 2023b), 7B and 13B, which are cutting-edge public foundation models whose internal representations are accessible. In our preliminary experiments, we have also tested on Llama, OPT (Zhang et al., 2022) and have similar observations as Llama-2. We report additional results with Mistral-7B-v0.1[2] based on experiments

[2]https://mistral.ai/news/announcing-mistral-7b/

Method	CoQA	TydiQA	TriviaQA		HotpotQA		Averaged
	0-shot	0-shot	0-shot	5-shot	0-shot	5-shot	0-shot
Llama-2-7B							
Pred. Entropy	0.715	0.590	0.697	0.768	0.650	0.669	0.663
LN-Pred. Entropy	0.725	0.621	0.678	0.756	0.631	0.725	0.664
Semantic Entropy	0.690	0.705	0.718	0.781	0.664	0.728	0.694
SAPLMA	0.666	0.628	0.624	0.641	0.536	0.599	0.614
P(True)	0.638	0.608	0.471	0.651	0.444	0.593	0.540
LID-MLE	0.758	0.735	0.754	0.761	0.701	**0.731**	0.737
LID-GeoMLE	**0.767**	**0.738**	**0.771**	**0.791**	**0.708**	0.729	**0.746**
Llama-2-13B							
Pred. Entropy	0.745	0.630	0.751	0.752	0.738	0.765	0.716
LN-Pred. Entropy	0.753	0.618	0.716	0.731	0.724	0.769	0.702
Semantic Entropy	0.758	0.740	0.736	0.786	0.708	**0.781**	0.736
SAPLMA	0.645	0.597	0.651	0.699	0.578	0.621	0.618
P(True)	0.649	0.624	0.511	0.662	0.518	0.581	0.576
LID-MLE	0.763	0.745	0.748	0.777	0.747	0.758	0.751
LID-GeoMLE	**0.772**	**0.759**	**0.775**	**0.793**	**0.749**	0.769	**0.764**

Table 1. Main results of predicting output correctness for four generative QA tasks. We compare our LID methods: LID-MLE, LID-GeoMLE with entropy-based and verbalized uncertainty estimation methods and a trained classifier. We show the results with Llama-2 7B and 13B. Results demonstrate the superior performance of our LID methods. The best scores are **bold**.

	TydiQA	CoQA
Pred. Entropy	0.630	0.712
LN-Pred. Entropy	0.619	0.723
Semantic Entropy	0.568	0.665
LID-MLE	0.722	0.752
LID-GeoMLE	0.734	0.761

Table 2. Results on Mistral-7B-v0.1. We only report TydiQA and CoQA results based on our additional experiments during the rebuttal period.

we have during the rebuttal period. Following the inference convention as in Touvron et al. (2023b), we conduct both zero-shot and few-shot inference. But for few-shot inference, we exclude the in-context examples when computing model representations. See inference examples with our prompt format in Appendix A.

Methods & Baselines For LID-MLE and LID-GeoMLE, we use 500 nearest neighbors when estimating LIDs for all datasets. We compare our methods with entropy-based uncertainty, verbalized uncertainty, and trained truthfulness classifiers on representations.

More specifically, for entropy-based ones, we consider predictive entropy (*Pred. Entropy*), length-normalized predictive entropy (*LN-Pred. Entropy*) (Malinin & Gales, 2020), and semantic entropy (*Semantic Entropy*) (Kuhn et al., 2022). Since precisely calculating the entropy is intractable over the infinite output space of generative QA tasks, we use Monte Carlo estimation of entropy on sampled outputs Malinin & Gales (2020). Formally, Pred. Entropy $= \frac{1}{N}\sum_{i=1}^{N} \log p\left(y_i|x\right)$, where y_i, $i = 1 \ldots N$ is the N sampled outputs. LN-Pred. Entropy simply replaces the log-likelihood with the length-normalized log-likelihood: LN-Pred. Entropy $= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{|y_i|} logp\left(y_i|x\right)$. Semantic Entropy groups outputs that are semantically equivalent to each other and calculate the entropy among groups: Semantic Entropy $= \frac{1}{|C|}\sum_{i=1}^{|C|} logp\left(C_i|x\right)$, where C_i is the summed likelihood of outputs in the i-th group. All entropy-based methods are sensitive to the choice of temperature in decoding and the number of sampled outputs. We follow Kuhn et al. (2022), and set the temperature to be 0.5 and the number of generated samples to be 10.

For verbalized uncertainty, we use $P\left(\text{True}\right)$ (Kadavath et al., 2022), which asks the model itself if its answer is correct. Then, take the probability for the model outputting the token 'True' as the truthfulness score.

For trained classifiers, we implement *SAPLMA* (Azaria & Mitchell, 2023) that train a multi-layer classifier on each dataset with 3,000 examples (1,500 truthful and 1,500 untruthful) to predict a binary truthfulness label for each example. The training setup follows Azaria & Mitchell (2023).

Evaluation Setup We use the area under the receiver operator characteristic curve (AUROC) to evaluate the effec-

tiveness of all baselines and our proposed LID method. The truthfulness prediction task is viewed as a binary classification task, and AUROC measures the performance under varying thresholds. The indicator function of truthfulness is given by $s(y_i, \hat{y}_i) = \mathbb{I}(\text{RougeL}(y_i, \hat{y}_i) \geq 0.5)$, following (Kuhn et al., 2022), where RougeL (Lin, 2004) score is a substring matching measurement commonly used to evaluate generative QA tasks. We show that the results is robust across different indicator function in Appendix B.

For TriviaQA and HotpotQA, we evaluate with both zero-shot and 5-shot in-context inference. We evaluate CoQA and TydiQA-GP with zero-shot as the context is long.

4.1. Sanity Check

To gain insights about the reliability of MLE-based estimators, we apply MLE-based methods on synthetic data with known ground truth dimensions, showing that they can approximately give the correct estimation. Additionally, we compare the intrinsic dimension obtained by MLE-based methods and other popular methods: KNN (Costa et al., 2005) and TwoNN (Facco et al., 2017a).

For synthetic data, we simulate two popular manifolds, namely a sphere and a norm, both having an original dimension of 4,096 and intrinsic dimensions of 10 and 20 (Table 3). We find that GeoMLE in general produces the most accurate estimation on those datasets, although MLE shows more negative bias. For real datasets, TwoNN and MLE-based methods give approximately similar estimations. Overall, MLE-based methods produce reasonable estimates in both scenarios. It is safe to assume that MLE offers a practical approximation for applications, especially when the true value of intrinsic dimension is not directly relevant to the application but rather comparisons of the values. Furthermore, note that MLE inherently estimates 'local' intrinsic dimension while other methods are 'global' estimators. Based on the points above, we adopt MLE-based methods in our study.

4.2. Main Results

The main results for Llama-2 are summarized in Table 1. The partial results for Mistral-7B-v0.1 are in Table 2. See Appendix C for the corresponding frequency histograms. Overall, we observe that LID-GeoMLE outperforms entropy-based methods by 0.05 points for 7B and 0.03 points for 13B on AUROC, while verbalized uncertainty are not comparable to the above methods. Notably, the performance of uncertainty-based methods is contingent on whether we present in-context examples to the LLMs while LID methods are more stable regarding this. For example, on TriviaQA with 0-shot, LID methods outperform the best entropy-based uncertainty estimation by 8%. The improvements comparing LID-MLE and LID-GeoMLE

Dataset	m	TwoNN	KNN	MLE	GeoMLE
Sphere	10	8.78	4	8.63	8.65
Sphere noise	10	13.97	4	11.45	9.64
Norm	20	17.54	9	15.54	20.33
Norm noise	20	17.81	2	15.72	22.36
TriviaQA	T	10.10	5	9.37	8.68
	F	15.45	10	12.43	11.97
CoQA	T	16.21	-	14.50	13.98
	F	17.79	-	16.62	16.10

Table 3. Sanity Check on intrinsic dimension estimate. We compare MLE with other famous estimates on synthetic data and our QA datasets. All with 1000 samples. For synthetic data, the column 'm' represents the ground-truth intrinsic dimensions. For simulated data, we report results on truthful (T) and untruthful (F) samples. KNN cannot provide reasonable estimates on CoQA.

	TriviaQA	HotpotQA	TydiQA	CoQA
TriviaQA	0.748	0.729	0.734	0.745
HotpotQA	0.724	0.747	0.726	0.724
TydiQA	0.765 ↑	0.744	0.745	0.736
CoQA	0.747	0.753 ↑	0.751 ↑	0.763

Table 4. Robustness to cross-task neighbors. Performance when the neighbors are from different datasets. The left column is the neighbors' dataset while the top column shows the tested dataset. The performance in general decreases slightly but is still effective.

suggest that a more sophisticated LID estimation method would bring additional benefits to the truthfulness prediction performance.

Moreover, notice that another representation-level method, SAPLMA, fails to match the performance of LID-based methods. This implies that LID-based methods remain powerful when truthful directions are hard to obtain. We show in Appendix D that dimension reduction methods like t-SNE also fail to find such directions. The findings in (Marks & Tegmark, 2023) might not hold in a practical scenario.

4.3. Robustness Study

We study the robustness of our proposed LID technique.

Robustness to hyperparameters n and T. We investigate how the performance and LIDs vary as a function of different numbers of dataset size n and neighbors T.

As illustrated in Figure 2, the performance on AUROC increases as we consider more neighbors, reaching its maximum as the number of neighbors used approaches the total dataset size. The performance of using LID method is comparable to or better than the best entropy-based uncertainty estimation methods even with around 200 neighbors.

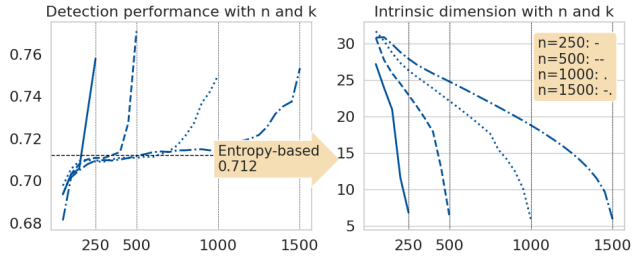

Figure 2. Robustness to n and T. Performance and intrinsic dimension as a function of the number of neighbors and total reference points. Both plots use the number of neighbors as X-axis. Different line styles indicate different numbers of reference points.

The intrinsic dimension shows a decreasing trend when we consider more neighbors, moving within the range of 6 to 30. The intrinsic dimension value will be lower with the same number of neighbors but with a larger dataset size. Again, MLE methods provide just approximations to the actual intrinsic dimension, which may contain bias but is of no direct concern in this application of detecting hallucinations.

Robustness to cross-task reference To understand how generalizable the LID feature is and whether we can use LIDs in the wild, we conduct experiments where the neighbors come from a different dataset. Results are shown in Table 4. We find that there is only a small decrease in performance in general, demonstrating the features used to estimate intrinsic dimension are generalizable to out-of-domain tasks. Also, notice that we observe some performance boosts when the reference samples come from a dataset that models have better performance on. For example, TydiQA reference samples improve the detection on HotpotQA. We leave further investigation to this observation to future work.

5. Analysis

In this section, we conduct a series of analyses on the characteristics of intrinsic dimensions of LLM representations. In the first two parts, we study how the intrinsic dimensions of model representations change across layers and during the autoregressive language modeling process. In the last part, we investigate how the effects of instruction tuning on intrinsic dimensions. Our study reveals that intrinsic dimension is indeed an insightful tool to understand LLMs.

5.1. The aggregated LIDs exhibit a hunchback shape in intermediate layers

To study the characteristics of the intrinsic dimensions insider the model layers, we aggregate individual LIDs and present the trends of the aggregated LIDs across different model layers. As depicted in Figure 3, across the four datasets, the intrinsic dimensions exhibit a *hunchback* shape, akin to the observation in the vision domain (Ansuini et al., 2019): the averaged LID values initially increase from the bottom to middle layers, and then decrease from the middle to the top layers. Note that intrinsic dimensions represent how many dimensions are needed to encode the information without significant loss. The observed hunchback phenomenon suggests that LLMs gradually capture the information in the context in the first few layers, and then condense them in the last layers to map to the vocabulary.

Furthermore, we observe a close relation between the intrinsic dimension values and the performance of predicting individual truthfulness. The two curves: the LID curve with dots (blue) and the detection performance curve (orange) show similar trends. Nevertheless, there is a 'shifting behind' effect: the variants in LID values are reflected one or two layers later in the prediction of truthfulness. We hypothesize that once the model encodes sufficient information, as indicated by the absolute LID values, additional transformations in later blocks are required to convert these encoded features into indicators of truthfulness. Empirical verification and further investigation of this phenomenon could be explored in future work.

5.2. The intrinsic dimensions are consistently lower for human answers at different positions

Next, to investigate how intrinsic dimensions vary when modeling truthful and untruthful outputs, we compare the LIDs of untruthful answers with the LIDs of their corresponding ground-truth answers at each position, using another set of complete correct answers as reference points. We use the TriviaQA dataset as an example. Figure 4 illustrates the aggregated results. We observe that the intrinsic dimensions of ground-truth are consistently lower than model generations at different positions. For ground-truth answers, a sharp decrease in the intrinsic dimensions happens when approaching the end of generations while this phenomenon does not hold for incorrect generations. This explains why selecting the last token gives the best performance when working with representations (Zou et al., 2023).

To conduct a controlled study, we construct a few examples where we explicitly mix human and model answers: we prompt the model with the first half of the correct answers but ask the model to generate the rest. As in Table 9, the mixed answers usually have higher LIDs compared to both human and model answers, even when the model answer is incorrect. This supports our hypothesis that incorrect answers mix more manifolds and thus have higher LIDs. More examples are displayed in Appendix E.

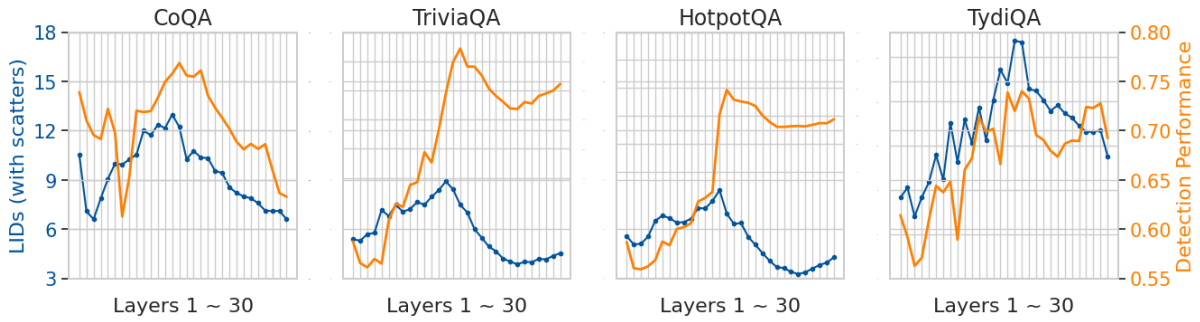

Figure 3. Plots for the aggregated LID values across model layers on the four QA datasets. The X-axis is the layer id, which is layer 1 to layer 30 for Llama-2-7B. The left Y-axis is the aggregated LID values and the right Y-axis is the detection performance (AUROC) values. The detection performance curve is in orange and the LID curve is in blue with markers. We show that there is a hunchback shape in the LID values across layers. The LID values closely correlate with the performance of detection and exhibit a 'shift behind' phenomenon.

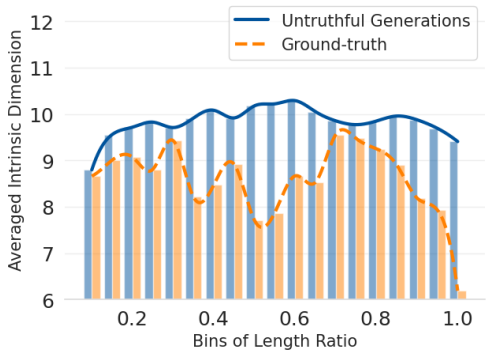

Figure 4. Bars for intrinsic dimensions of ground-truth answers (orange dashed line) and untruthful model generations (blue line) as the language modeling proceeds. The X-axis represents buckets of different ratios of the total lengths.

`Answer these questions: Q: Who played the title roll in the Flint films? A:`
- **Model output:** `John Cusack` [7.54, 8.61, 13.34, **5.05**]
- **Ground-truth:** `James Coburn` [7.88, 9.96, **4.79**]
- **Mixed output:** `James Garner` [7.87, 8.72, 4.58 , **6.98**]

Table 5. Examples of mixing distributions increases LIDs. The Blue part is the model continuation for the ground-truth. The numbers in the list represent LID value for each position.

5.3. The intrinsic dimensions increase while instruction tuning and correlate with model performance

Finally, we study how instruction tuning affects LLMs' intrinsic dimensions. Instruction tuning adapts a pre-trained LLM to solve diverse tasks by training LLMs to follow declarative human instructions (Wang et al., 2022; Taori et al., 2023). We follow this paradigm and investigate the representation LIDs.

Experimental Setup We use the SUPER-NI (Wang et al., 2022) for training, which contains 756 training tasks and 200 examples for each training task. We fine-tune a Llama-2-7B model for 3,000 steps, roughly 3 epochs, on SUPER-NI's training set. We track every 300 steps during the tuning process and evaluate those checkpoints for their accuracy on TydiQA and TriviaQA, as well as the intrinsic dimension footprints. For both TydiQA and TriviaQA, we randomly sample 1,000 test examples and repeat the experiments three times. We use T=500 nearest neighbors for estimating the LIDs on TriviaQA and TydiQA.

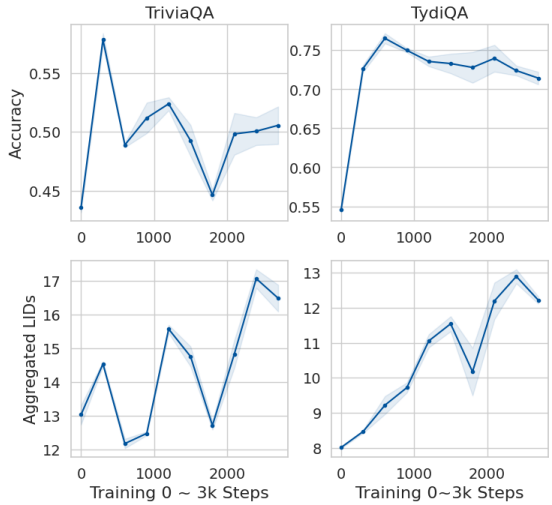

Figure 5. Plots for the accuracy and intrinsic dimension on TriviaQA and TydiQA during instruction tuning. The X-axis is the training steps. We train 3,000 steps in total and show checkpoints every 300 steps. The Y-axis is the performance for the top two figures and the aggregated LID values for the bottom two figures.

The intrinsic dimension grows while training for a longer time As illustrated in Figure 5, instruction tuning brings a performance boost for both TriviaQA and TydiQA, while the boost is more significant for TydiQA. We also find that the

aggregated LID values show an increasing trend along with the training process, although there are more fluctuations for TriviaQA than TydiQA. During instruction tuning, models are tuned on diverse tasks and composed distributions. The increase in intrinsic dimensions of LLM representations possibly implies that they become richer.

The aggregated LID values correctly predict fluctuations in generalization ability during training We observe that the performance on both TriviaQA and TydiQA reaches a local minimum at some intermediate checkpoints during instruction tuning, while the intrinsic dimensions decrease correspondingly. For example, at step 600 and step 1800, the performance for TriviaQA fluctuates and reaches the local minimum. This is reflected in the curve of LID values, where at step 600 and step 1800, the LID values are the lowest locally. Similar observations exist for TydiQA at step 1800. This suggests that one may use the intrinsic dimension as a signal to select model checkpoints.

6. Discussion

In this section, we want to discuss some limitations of LID. First, as model representations are trained with data from the human distribution, the truthfulness LID characterizes is mainly based on the world knowledge, i.e., it is distinguishing whether models are generating content consistent with world knowledge or not. The performance of LID to detect other properties of model generations besides truthfulness is yet to be explored. For example, if the generated content is faithful to some context, or if it is adversarial or harmful.

Second, the performance of LID on different formats of inputs and outputs like multi-modal and code-switching inputs and document-level outputs, is not explored in this work.

7. Conclusion

In this paper, we proposed to use LIDs to characterize and predict the correctness of LLM outputs, which achieved better performance compared to prior methods. We showed several empirical observations about model intrinsic dimensions, including the variants of them with model layers, autoregressive language modeling, and the effects of instruction tuning. This opened up a new direction to consider quantifying model truthfulness for future work.

Acknowledgement

We thank UCLA-NLP and UCLA-PlusLab members, as well as anonymous reviewers, for their invaluable feedback while preparing this draft. We thank Po-Nien Kung for providing the codebase of instruction tuning. This project is partially supported by a sponsored research award by Cisco Research, and an ONR grant N00014-23-1-2780.

Impact Statement

This paper discusses an important step for the safe deployment of LLMs. LLMs have demonstrated their remarkable capabilities, but human trust on LLMs are still limited because of their tendency to generate hallucinations secretly. We propose to quantify and characterize model hallucinations through intrinsic dimensions of model intermediate representations. On the one hand, this method itself helps people abstain from trusting incorrect generations. On the other hand, it can serve as the backbone for some hallucination mitigation methods.

This paper mainly discusses LLM generations in English. Results may be biased towards the English-speaking population. However, adapting this LID method to other languages does not require much additional effort. Moreover, the paper focuses primarily on question-answering scenarios. We encourage future work to implement this method on diverse tasks and languages, as well as cross-lingual settings.

References

Amsaleg, L., Chelly, O., Furon, T., Girard, S., Houle, M. E., Kawarabayashi, K.-i., and Nett, M. Estimating local intrinsic dimensionality. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 29–38, 2015.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Azaria, A. and Mitchell, T. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*, 2023.

Birdal, T., Lou, A., Guibas, L. J., and Simsekli, U. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34:6776–6789, 2021.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2022.

Campadelli, P., Casiraghi, E., Ceruti, C., Rozza, A., et al. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015:1–21, 2015.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

Costa, J. A., Girotra, A., and Hero, A. O. Estimating local intrinsic dimension with k-nearest neighbor graphs. *IEEE/SP 13th Workshop on Statistical Signal Processing, 2005*, pp. 417–422, 2005. URL https://api.semanticscholar.org/CorpusID:15177727.

Duan, J., Cheng, H., Wang, S., Wang, C., Zavalny, A., Xu, R., Kailkhura, B., and Xu, K. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*, 2023.

Facco, E., d'Errico, M., Rodriguez, A., and Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7, 2017a. URL https://api.semanticscholar.org/CorpusID:3991422.

Facco, E., d'Errico, M., Rodriguez, A., and Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017b.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Gomtsyan, M., Mokrov, N., Panov, M., and Yanovich, Y. Geometry-aware maximum likelihood estimation of intrinsic dimension. In Lee, W. S. and Suzuki, T. (eds.), *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pp. 1126–1141. PMLR, 17–19 Nov 2019. URL https://proceedings.mlr.press/v101/gomtsyan19a.html.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., and Matias, Y. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3905–3920, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know, 2022.

Kamath, A., Jia, R., and Liang, P. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, 2020.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2022.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Levina, E. and Bickel, P. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Lin, Z., Trivedi, S., and Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018.

Malinin, A. and Gales, M. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2020.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2020.

Reddy, S., Chen, D., and Manning, C. D. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Ren, J., Luo, J., Zhao, Y., Krishna, K., Saleh, M., Lakshminarayanan, B., and Liu, P. J. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Piontkovskaya, I., Nikolenko, S., and Burnaev, E. Intrinsic dimension estimation for robust detection of ai-generated texts. *arXiv preprint arXiv:2306.04723*, 2023.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*, 2022.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhou, K., Jurafsky, D., and Hashimoto, T. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*, 2023.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Dataset and Inference Details

For datasets without context (HotpotQA and TriviaQA), we use the following textual input as prompts:

Answer these questions: \n Q: [question] \n A:

For datasets with context (TydiQA-GP and CoQA), we have the following template for prompts:

Answer these questions based on the context:\n Context: [a passage or a paragraph] \n Question: [question] Answer:

Table 6 shows some examples from those datasets with our inference format.

TriviaQA
- Answer the following questions: \n Q: The Great Barrier Reef is located off the coast of which Australian state? \n A: - Answer the following questions: \n Q: Whose 2006 album ""Back to Black"" had 6 Grammy Award nominations and five wins, tying the record for the most wins by a female artist in a single night, and made her the first British singer to win 5 Grammys?\n A: - Answer the following questions: \n Q: Who directed The Cable Guy? \n A:
HotpotQA
- Answer these questions: \n Q: Which film was created first, Tex or Miracle? \n A: - Answer these questions: \n Q: When was the company, of which Ernest Walter Hives was one time chairman, founded ? \n A:
CoQA
- Answer these questions based on the context: \n Context: Annie was helping her little brother Max pick flowers from the garden. They wanted to put the flowers in a jar to put on the kitchen table. Mother's Day was the next day and their mother loved fresh flowers. After they picked flowers and put them in a jar, Max asked Annie if they could have a snack. Annie took Max into the kitchen and got out an apple to slice up. They sat down at the table looking at the flowers and ate their apple slices. There was a window in the kitchen that let in sunlight. ""Hey!"" Max said, pointing at one of the roses in the jar. ""There's something moving on that rose."" Annie looked more closely at the flowers. ""It's a ladybug,"" she said. ""We need to take it back outside."" Suddenly the ladybug began flying around the kitchen. Max jumped up and ran around trying to catch it. At last he clapped his hands around it. ""Careful!"" said Annie. Max walked outside and let the ladybug go. \n Question: Was Max an older brother? \n Answer:
TydiQA-GP
- Answer the following question based on the information in the given passage: \n Passage: The Kingdom of Aksum (also known as the Kingdom of Axum, or the Aksumite Empire) was an ancient kingdom located in what is now Tigray Region (northern Ethiopia) and Eritrea.[2][3]. Axumite Emperors were powerful sovereigns, styling themselves King of kings, king of Aksum, Himyar, Raydan, Saba, Salhen, Tsiyamo, Beja and of Kush.[4] Ruled by the Aksumites, it existed from approximately 100 AD to 940 AD. The polity was centered in the city of Axum and grew from the proto-Aksumite Iron Age period around the 4th century BC to achieve prominence by the 1st century AD. Aksum became a major player on the commercial route between the Roman Empire and Ancient India. The Aksumite rulers facilitated trade by minting their own Aksumite currency, with the state establishing its hegemony over the declining Kingdom of Kush. It also regularly entered the politics of the kingdoms on the Arabian Peninsula and eventually extended its rule over the region with the conquest of the Himyarite Kingdom. The Manichaei prophet Mani (died 274 AD) regarded Axum as one of the four great powers of his time, the others being Persia, Rome, and China.[2][5] \n Question: When did the Kingdom of Aksum end? \n Answer:

Table 6. Examples of datasets.

	TriviaQA	HotpotQA	TydiQA	CoQA
Llama-2-7B	0.66	0.26	0.44	0.57
Llama-2-13B	0.72	0.33	0.50	0.62

Table 7. Accuracy of Llama-2-7B and Llama-2-13B on the four datasets with $s\left(y_i, \hat{y}_i\right) = \mathbb{I}\left(\text{RougeL}\left(y_i, \hat{y}_i\right) \geq 0.5\right)$

The performance with LLama-2 is shown in Table 7, which roughly matches the publicly reported performance.

B. Different Indicator Functions

We evaluate the sensitivity of our methods towards different indicator functions. We consider two new indicator functions:

1) $s(y_i, \hat{y}_i) = \mathbb{I}(\text{RougeL}(y_i, \hat{y}_i) \geq 0.3)$, i.e., changing the thresholds in the Rouge-L metric from 0.5 to 0.3;

2) $s(y_i, \hat{y}_i) = \mathbb{I}(\text{NLI}(y_i, \hat{y}_i) == \text{entailment})$. We use a natural language inference (NLI) model to judge the semantic similarity of two answers as the indicator function. The model is tuned for judging whether a premise semantically entails a hypothesis, where the ground-truth answer is used as the premise and the model-generated answer is used as the hypothesis. The NLI model is based on T5-XXL from Honovich et al. (2022).

We evaluate with TriviaQA and CoQA on Llama-2-7B. The only difference with our main results is that the indicator functions are changed to the above two functions. Results are shown in Table 8. We show that despite of small performance variants, LID-MLE and LID-GeoMLE outperform the best uncertainty-based methods across different indicator functions.

Method	CoQA	TriviaQA 0-shot	TriviaQA 5-shot
Rouge-L ≥ 0.3			
Semantic Entropy	0.683	0.726	0.791
SAR (Duan et al., 2023)	0.694	0.731	0.795
LID-MLE	0.744	0.761	0.783
LID-GeoMLE	**0.756**	**0.784**	**0.801**
NLI Entailment			
Semantic Entropy	0.666	0.715	0.779
SAR (Duan et al., 2023)	0.671	0.758	0.783
LID-MLE	0.723	0.762	0.783
LID-GeoMLE	**0.735**	**0.781**	**0.790**

Table 8. Detect incorrect answers for Llama-2-7B on TriviaQA and CoQA. We use two other indicator functions based on Rouge-L threshold of 0.3 and NLI entailment. Results demonstrate that LID methods is robust to different indicator functions.

C. Frequency Histogram

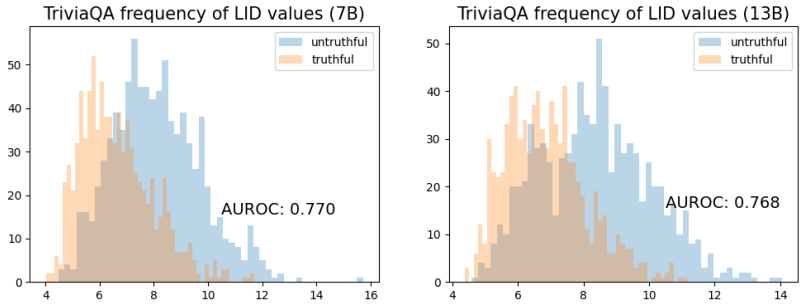

Figure 6. Frequencies of the LID values for truthful and untruthful data on TriviaQA, with Llama-2-7B (left) and Llama-2-13B (right). X-axis is the LID values while Y-axis is the number of occurs.

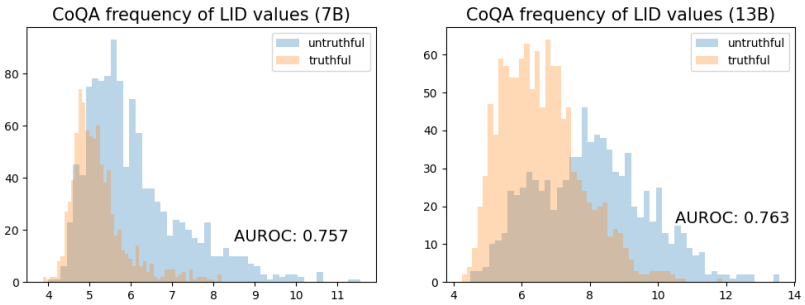

Figure 7. Frequencies of the LID values for truthful and untruthful data on CoQA, with Llama-2-7B (left) and Llama-2-13B (right). X-axis is the LID values while Y-axis is the number of occurs.

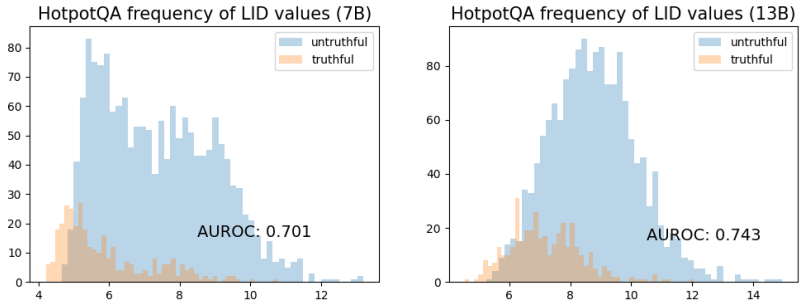

Figure 8. Frequencies of the LID values for truthful and untruthful data on HotpotQA, with Llama-2-7B (left) and Llama-2-13B (right). X-axis is the LID values while Y-axis is the number of occurs.

D. Visualization of truthful and untruthful answers with t-SNE

We show the t-SNE scatters that reduce the dimensions of original representations into a two-dimensional space. On CoQA and TydiQA, there are vague clusters of truthful and untruthful generations. But on TriviaQA and HotpotQA, there is no clear cluster. Overall, the visualization shows that dimension reduction methods like t-SNE fail to distinguish truthful answers from untruthful answers.

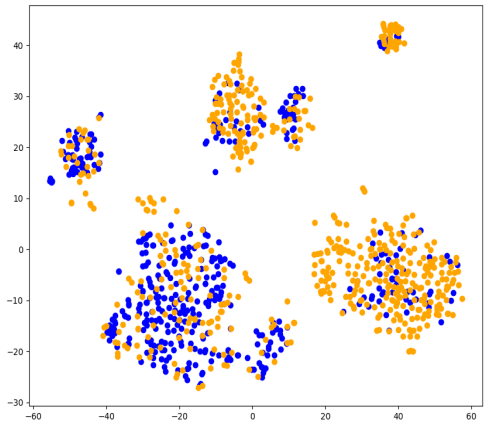

Figure 9. t-SNE on CoQA.

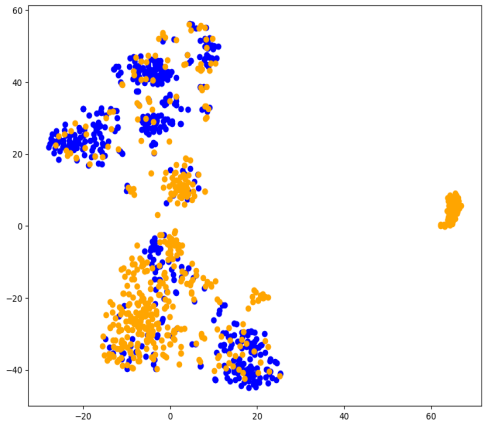

Figure 10. t-SNE on TydiQA.

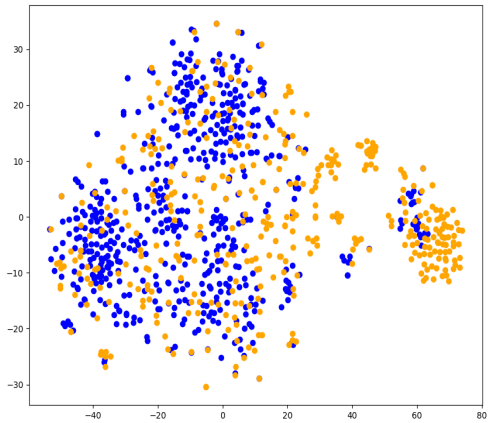

Figure 11. t-SNE on TriviaQA.

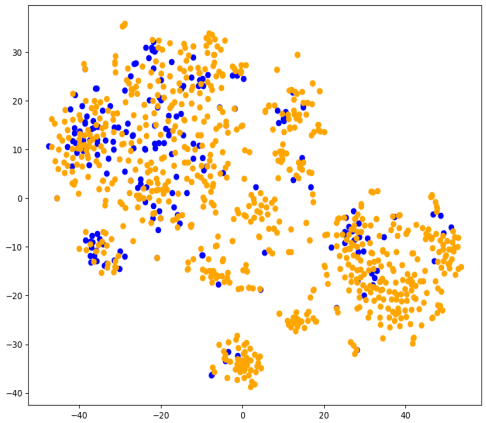

Figure 12. t-SNE on HotpotQA.

E. Examples for the controlled mixing data

We show more examples or our synthetic experiments where we ask models to continue a ground-truth answer,

`Answer these questions: Q: Who played the title roll in the Flint films? A:`
- **Model output:** `John Cusack` [7.54, 8.61, 13.34, 5.05] - **Ground-truth:** `James Coburn` [7.88, 9.96, 4.79] - **Mixed output:** `James Garner` [7.87, 8.72, 4.58 , 6.98]
`Answer these questions: Q: Which man other than Billy McNeill and Martin O'Neil has managed both Aston Villa and Celtic? A:`
- **Model output:** `1970s Celtic manager Jock Stein.` [7.98, 8.05, 8.98, 7.08, 8.76, 10.15, 8.22, 11.30, 9.22, 6.78, 6.76, 8.57, 7.33, 6.87] - **Ground-truth:** `DR JOSEF VENGLOS` [8.32, 7.66, 8.11, 8.87, 8.80, 7.74, 8.38, 10.76, 5.74] - **Mixed output:** `DR JOSEF SMEDA` [8.32, 7.66, 8.11, 8.87, 8.80, 7.50, 11.17, 7.46]
`Answer these questions: Q: Which famous vehicle has used license plates SCV 00919 and SCV 1 (through to SCV 9)? A:`
- **Model output:** `007's Aston Martin DB5` [9.50, 10.02, 11.33, 9.06, 8.82, 10.25, 8.07, 9.75, 6.79, 8.46, 6.70] - **Ground-truth:** `The Popemobile` [9.00, 9.23, 9.53, 6.72] - **Mixed output:** `The Black Pearl` [9.00, 8.52, 11.39, 7.04, 7.86]
`Answer these questions: Q: In the J. M. Barrie play who was the resourceful butler to the Earl of Loam? A:`
- **Model output:** `Jeeves` [8.75, 10.51, 6.43] - **Ground-truth:** `THE ADMIRABLE CRICHTON` [6.74, 7.82, 9.03, 9.86, 9.91, 10.19, 10.02, 12.58, 5.41] - **Mixed output:** `THE ADMIRABLE BERTIEl` [[6.74, 7.82, 9.03, 9.86, 9.91, 7.39, 8.11, 7.29]
`Answer these questions: Q: With what invention do you associate the name of Mr. Whitcomb Judson? A:`
- **Model output:** `The sewing machine` [9.46, 9.75, 9.40, 5.55, 7.49] - **Ground-truth:** `ZIP fastener` [8.69, 10.32, 11.12, 6.47] - **Mixed output:** `ZIP PERl` [[8.69, 10.32, 8.37, 9.05]

Table 9. Examples of mixing distributions increases LIDs.