On the Structural Inevitability of LLM Hallucinations

Dhruv Patankar[1], Sachit Ramesha Gowda[2]

[1]Shunya Research , dhruv@shunyaresearch.systems
[2]Shunya Research , sachit@shunyaresearch.systems

March 16, 2026

Abstract

Hallucination in large language models is widely treated as an engineering defect, addressed through scaling, better data, or alignment techniques. This paper argues that hallucination is not a defect but a structural consequence of applying function approximation to domains with inconsistent target mappings. Drawing on the Universal Approximation Theorem, we show that LLMs succeed where training data is consistent (grammar, syntax) and fail where it is not (facts, open-domain conversation). When a model is trained on contradictory sources, the optimal solution under negative log-likelihood is an interpolation of conflicting distributions, not the ground truth. Conversational context compounds this by expanding the input space with each turn, ensuring the model operates perpetually in extrapolation regimes with no mechanism to detect this. We further argue that the core failure is not extrapolation itself but the model's blindness to it, as no uncertainty signal is embedded in the training objective. Scaling and RLHF do not resolve this: they smooth the extrapolation without changing the underlying geometry. A viable solution requires architectural changes that give models the ability to estimate their distance from the training manifold at inference time.

1 Introduction

Hallucination in large language models is routinely described as a problem to be solved. The proposed solutions follow a familiar pattern: more data, larger models, better alignment, retrieval augmentation. Implicit in all of these is a shared assumption: that hallucination is an engineering defect, correctable in principle by sufficient effort in the right direction. This paper challenges that assumption.

We argue that hallucination in conversational LLMs is not a defect. It is a structural consequence of applying a static training objective to a non-stationary, high-entropy, unbounded input process. It cannot be fixed by scaling.

It cannot be fixed by better data. Any solution that does not address the geometric and information-theoretic root cause will at best suppress symptoms while leaving the underlying condition intact.

The central argument of the paper is as follows. Universal Approximation Theorem (UAT), as explained in Cybenko (1989) states that any continuous function can be approximated, to a desired degree of accuracy by a neural network, given enough layers. LLMs are concrete embodiments of UAT, according to Wang and Li (2024). For this reason, LLMs succeed in cases where the target mapping is consistent, such as grammar. However LLMs fail sometimes in domains where the target mapping is not consistent. The most common example of this is factual truth, where different sources may contain different information, causing hallucination.

2 Background

2.1 LLMs as Universal Function Approximators

A function in mathematics, assigns each element from a set X to exactly one element in another set Y, where the set X is called the domain and the set Y is called the codomain.

Sparse transformers are universal approximators of continuous, fixed length, sequence-to-sequence functions on compact domains, provided they meet the conditions of appropriate sparsity/connectivity, and a suitable attention probability map. The key caveat here is that this behaviour is only defined for a fixed domain, as defined above. It has no bearing on the behavior outside this domain. The actual learning process of the model is approximating $P(\text{next token} \mid \text{context})$, which is a mapping from the context to the probability distribution. Since every context maps to a single probability distribution, this is a function. Therefore hallucination is not a sign of something being broken, the model is doing exactly what it was designed to do.

2.2 What LLMs actually optimize

LLMs use NLL(Negative Log Likelihood) in their training process. The objective of the NLL process is to maximize the likelihood of observing data, given some parametric conditions. The specific objective in LLM training is to minimize the negative log likelihood of $P(\text{next token} \mid \text{context})$, as discussed above. The LLM is rewarded for producing statistically correct continuations.

However, in this training process there is no term to check factual correctness, or distance from training distribution. Both a fluent true statement and a fluent falsehood will receive a similar weight update from the gradient descent if they are statistically similar to the training data.

3 The Consistency Requirement

According to UAT, function approximation works well when the target mapping is consistent and stable across the training distribution. Grammar and syntax have a consistent one to one mapping, there are no contradictions about whether sentences are grammatically correct or not. Syntactic well-formedness is also rule based, as shown by Chomsky (1957). Hence, the transformer is learning a system with consistent rules.

Structure and causality are similar. The training corpus does not contradict itself about whether 'if X then Y' is valid. Therefore the statistical distribution of the grammatical continuations align closely with the grammar output that the human wants. However, there is a distinction to draw here. It is not that grammar is 'mathematical' but rather, the training corpus is consistent about grammar so the function that is approximated is well defined.

Let $\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$ be a corpus of n documents. For a given context $x_{1:t}$, let $P_i(x_{t+1} \mid x_{1:t})$ denote the empirical continuation distribution induced by document Di — that is, the distribution over tokens that follow $x_{1:t}$ in Di.

A context $x_{1:t}$ is consistent with respect to corpus $\mathcal{D}$ if the continuation distributions across all documents that contain $x_{1:t}$ are in agreement. Formally, for all i, j such that $x_{1:t} \in D_i$ and $x_{1:t} \in D_j$:

$$D_{\mathrm{KL}}(P_i(x_{t+1} \mid x_{1:t}) \| P_j(x_{t+1} \mid x_{1:t})) \approx 0$$

A domain is consistent if this condition holds for all contexts within it.

For grammatical context, $x_{1:t}$, the set of valid continuations, $\mathcal{G}(x_{1:t})$, remains stable throughout the corpus. If you sample two documents from P_{data}, that have a common grammatical context, they will draw their continuations from the same distribution. The sources do not contradict each other.

We can express this as a low-variance condition. Let $P_i(x_{t+1} \mid x_{1:t})$ be the distribution over continuations in document i. The consistency condition is:

$\mathrm{Var}[P_i(x_{t+1} \mid x_{1:t})] \approx 0$

The variance is low in the case of grammar because all sources follow the same rules. The model approximates a stable target and function approximation works well because of this.

Now let's examine factual content. For a question like, "Is coffee good for you?", different sources in the corpus give different answers. As a result, the variance is high:

$\mathrm{Var}[P_i(x_{t+1} \mid x_{1:t})] \gg 0$

When the model minimizes $L(\theta)$ over inconsistent data, the optimal solution is the average of contradictory distributions, not the true answer. The model learns something like:

$P(x_{t+1} \mid x_{1:t}) = \mathbb{E}_i[P_i(x_{t+1} \mid x_{1:t})]$

Finding the expected value of these contradictory sources is what produces hallucination. The model correctly finds the optimal solution for the given objective. But this is not the truth, this is a confident-sounding interpolation of contradictory claims.

Facts have grounding in the real world. Linguistically, there is no distinction between facts and fiction. The same well formed question can have multiple different answers, depending on the context of the situation. As a result, the training corpus is inconsistent about facts, sources may contradict one another and information may change over time. When a model receives contradictory inputs from multiple sources, it does not try to resolve the contradiction, instead it averages the inputs from the sources. This leads to true sounding outputs with no grounding in reality.

Essentially, the ground truth mapping from question to correct answer is not well defined. It is context dependent, time dependent and contradicts across sources. The model approximates distribution of claims, not the truth.

4 Unbounded Context and Extrapolation Blindness

4.1 How Conversation Amplifies the Problem

Most tasks performed by the LLM have bounded input distributions. Images are constrained by the laws of physics, code is constrained by syntax, translation is constrained by grammar. Conversational context has no such bound. It is generated by the user, turn dependent, topic-shifting and is not constrained by any physical or syntactic prior. With every additional character added to the input, the input distribution changes. As a result, there is no stable $P(\mathrm{x})$ to approximate, Lee et al. (2025).

In sentence completion, more context narrows the output distribution. This is not true for conversational context, as it does not converge to a low entropy regime. At any context length, the conversation may diverge, and multiple distinct responses may be valid at any context length. As a result, the model is always extrapolating, with no way to check its distance from its training distribution.

In bounded tasks, the input space, X, is fixed and the model converges to a low-entropy region as the context increases. In conversation however, the input space at a turn,t, is:

$X_t = X_{t-1} \times V^*$

Where V^* is the space of all possible utterances. With each turn the context space expands instead of narrowing. The manifold that has to be covered by the model grows as t grows. Hence the probability mass is spread over a larger space and entropy does not converge. This directly links to the extrapolation argument, Balestriero et al. (2021), as the dimension d^* of the context vector grows with each turn, which means the exponential data requirement for interpolation gets worse as the conversation size increases.

4.2 The Blindness Problem

The previous sections seem to suggest that the problem is extrapolation. While extrapolation is an issue, it is the unawareness of extrapolation which is a critical issue. If the model was aware of its extrapolation, it would not be a problem. The critical issue is that the model has no way of knowing if it is extrapolating. The objective of the NLL in the LLM training process is to minimize $P(\text{next token} \mid \text{context})$. The weight update in the gradient does not contain a term that tells the model the geometric position relative to the manifold.

A hallucination receives the same training signal as a true statement if they are both statistically plausible in the training corpus. This could be solved by having some sort of uncertainty quantification methods, in the training process. However, most current uncertainty quantification methods are all external and post-hoc, as shown by Shorinwa et al. (2025), making the above impossible. As a result, the model produces fluent, confident text, agnostic of the distance from its training distribution.

5 The Current Approach and Future Solutions

5.1 Why scaling does not fix this

The obvious objection to these concerns is that scaling will solve these issues. With bigger models and more data, hallucination becomes increasingly less common. However, more data does not resolve inconsistencies in the training corpus. In fact, it may make it worse, by introducing more contradictory sources. Larger models simply mean the model scales, by having more parameters and layers. It does not mean the geometry of the model changes. With larger models, the extrapolation process is smoother, but the models are still extrapolating.

In high dimensional settings, interpolation almost surely does not happen, so conversational outputs at inference time are effectively extrapolations, as shown by Balestriero et al. (2021). For interpolation to hold, the dataset must grow exponentially with embedding dimension. This requirement is not possible for conversational LLMs. Another proposed solution is reinforcement learning with human feedback (RLHF) Ouyang et al. (2022). While RLHF improves surface level behavior and reduces hallucinations on some benchmarks, it still does not incorporate any uncertainty measurement into the generation process. In many cases, RLHF causes sycophantic behavior in models, driven by human preference that prefers a yes-man to a truthful response (Sharma et al., 2025). Retrieval-augment generation (RAG) tends to be more helpful, because it externalizes the world knowledge, which is more inconsistent, to an actual lookup. However, this is still an architectural workaround and does not address the underlying shortcomings.

5.2 What a Solution Would Require

Correctly diagnosing the problem is a prerequisite for finding any viable solution. The primary requirement is that the model should have a mechanism to estimate its distance from the training manifold at the time of inference. This would help the model be aware of its extrapolation. This is not a problem that can be solved by better prompt engineering, it would require complete architectural or training changes.

Some possible directions this solution could take are:

- Training time uncertainty objectives that penalize confident sounding generation in extrapolation.
- Manifold distance measurement during inference-time
- Separating retrieving facts from generative continuation so that extrapolation distance can be measured for individual claims.

Semantic entropy, Farquhar et al. (2024) is an existing approach, but it still remains external to the model and serves as a diagnostic tool rather than as a remedy. A viable solution will have to be internal.

6 Limitations

There are a few limitations of this paper that we must address. The first one is that our framing of consistency is a simplification. We treat domains as either consistent or inconsistent. In reality, consistency may be a spectrum. Some factual domains may be highly consistent, like arithmetic, and some grammatical domains may be less consistent, like variations in dialects of the same language.

The second limitation is that the arguments in the paper are structural, not quantitative. We do not measure the consistency across domains or entropy of conversational context empirically and we do not quantify the extrapolation distance for any model. Our arguments remain more theoretical.

The third limitation is about the limits of UAT grounding. The results of UAT, Yun et al. (2020), apply to sparse transformers on compact domains. Modern LLMs differ architecturally in a way that may not be covered by the theorem (scale, dense attention, positional encoding). We assume that the UAT framing generalizes, but this has not been proven formally yet.

7 Conclusion

Through this paper we have established that hallucination is a predictable consequence of applying function approximation to a domain with inconsistent mapping. Grammar works because the training data is consistent about grammar. Facts fail because the training data is not consistent about facts, and facts are dependent on the ground truth in a way that our current training methods

cannot solve. Conversation is the hardest case of this, as it adds inconsistent, unbounded context on top of the inconsistency problem. To find a solution, we must correctly frame it as a geometry and consistency problem rather than something that can be solved solely by data and scaling. Framing hallucination as a geometry and consistency problem and not as a data quality problem transforms our solution space. It converts an unbounded, empirical search for more and better data into a well defined requirement for better architecture. It allows us to build models which are self aware, in the context that they can reason about their distance from their own training data.

References

Randall Balestriero, Jérôme Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485*, 2021.

Noam Chomsky. *Syntactic Structures*. Mouton, The Hague, 1957.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL `https://doi.org/10.1038/s41586-024-07421-0`.

Jing Yang Lee, Kong Aik Lee, and Woon-Seng Gan. Modeling the one-to-many property in open-domain dialogue with LLMs. In Ofir Arviv, Miruna Clinciu, Kaustubh Dhole, Rotem Dror, Sebastian Gehrmann, Eliya Habba, Itay Itzhak, Simon Mille, Yotam Perlitz, Enrico Santus, João Sedoc, Michal Shmueli Scheuer, Gabriel Stanovsky, and Oyvind Tafjord, editors, *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 276–290, Vienna, Austria and virtual meeting, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-261-9. URL `https://aclanthology.org/2025.gem-1.24/`.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL `https://arxiv.org/abs/2203.02155`.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda

Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025. URL https://arxiv.org/abs/2310.13548.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions, 2025. URL https://arxiv.org/abs/2412.05563.

Wei Wang and Qing Li. Dynamic universal approximation theory: The basic theory for transformer-based large language models, 2024. URL https://arxiv.org/abs/2407.00958.

Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. $o(n)$ connections are expressive enough: Universal approximability of sparse transformers, 2020. URL https://arxiv.org/abs/2006.04862.