

Entropy evolution in multilingual transformer models : An information-theoretic perspective

Arti Bansode *

Anita Chaware †

Department of Computer Science

SNDT Women's University

Juhu Campus

Mumbai 400049

Maharashtra

India

Abstract

Entropy is used to quantify uncertainty in probabilistic representations. Hence, studying how uncertainty evolves across layers in LLMs can help us understand how these models process different languages. This paper presents a layer-wise entropy analysis of how information is processed across layers of four different Transformer models using parallel data in English, Hindi, Marathi, and Gujarati. The Shannon entropy is computed for each layer and several information-theoretic metrics are derived, including entropy change, CR, JSD and LDI. In addition, four scalar measures like NLC, ACD, ES, and CLAS are used to summarise overall compression behaviour. The results show that autoregressive models exhibit largely monotonic entropy compression, whereas the multilingual model shows oscillatory compression and expansion patterns. English undergoes stronger and more efficient entropy reduction than Indic Languages, while Hindi and Marathi show closely aligned entropy trajectories. These findings support the use of entropy-based analysis as a diagnostic tool for analysing architectural behaviour and cross-lingual representational disparities in transformer models.

Subject Classification (2020): 68T07, 94A17, 68T50, 62B10.

Keywords: Entropy, LLMs, Information theory, Entropy analysis.

1. Introduction

The discovery of Large Language Models (LLMs) has transformed the field of Natural Language Processing. However, interpreting how these models process and transform information across their layers needs to be studied. Standard evaluation metrics like accuracy & perplexity measure overall

* E-mail: artibansodephd@gmail.com (Corresponding Author)

† E-mail: anita.chaware@computersc.sndt.ac.in

performance but obscure the internal dynamics of uncertainty propagation within the model. Shannon entropy [1] H of a probability distribution P is defined as

$$H = -\sum_i p_i \log_2 p_i \quad (1)$$

where p_i are the probabilities of outcomes. It captures how “spread out” or uncertain a distribution is. In language models, entropy can be used to measure how much uncertainty remains about the next-token prediction or how compactly the model represents its hidden states. Prior work in information theory and deep networks has explored related ideas. The *information bottleneck* principal frames deep layers as successively compressing information [2]. Some studies have applied entropy or mutual information measures to neural networks [3], [4], [5]. However, most such analyses focus on English or do not directly address multilingual settings. In particular, the rich orthography and tokenization of Indic languages present unique challenges that have not been systematically investigated through an information-theoretic lens. From an information-theoretic perspective, we therefore ask: How does entropy evolve layer by layer in transformer models, and how do these dynamics differ across languages and architectures? To answer these questions, we introduce a layer-wise entropy framework for studying the model’s output distribution at each layer and for each language input. We further compute Shannon entropy along with several metrics to quantify compression and divergence. We study four transformer models with an input of 100 sentences in four languages. The main contributions of this study are: 1. We conduct the first multilingual, layer-wise entropy analysis of transformer LLMs across Indic languages. 2. We define and compute a set of information-theoretic metrics like layer wise entropy H , entropy change ΔH , CR, JSD and LDI to quantify the uncertainty dynamics. 3. We report empirical findings on our balanced dataset that reveal consistent compression–expansion patterns and language-specific entropy differences. We discuss the implications of these findings for model interpretability, pruning, and multilingual representation learning.

2. Related Works

Information Theory has been used several times for studying deep neural network behaviour. Entropy has also been used in advanced dynamical and hyper fuzzy systems to describe how information evolves and becomes more structured over time [6]. It has also been used for understanding compression in neural nets [2]. This approach is now being used for studying Transformer models. Ali et al. [7] introduced *Entropy-Lens*, a model-agnostic tool that computes Shannon entropy across the layers of pretrained transformers. Entropy-Lens demonstrates that layer wise entropy trajectories can distinguish different model families and correlate with task characteristics, providing a direct precedent for our layer wise entropy diagnostics. Similarly, layer-wise batch entropy has been proposed as a pruning signal; Danelli [8] showed that layers with persistently low entropy can be pruned with minimal accuracy loss, indicating that such layers carry less “new” information. Other related work has

studied attention mechanisms and activations: for instance, Queipo-de-Llano et al. [9] analyse "compression valleys" in decoder layers and link these to attention patterns. Though these works focus on English text and large models, they support the interpretation that drops in intermediate-layer entropy correspond to structured computation phases. More broadly, many studies use layer-wise probing to show that linguistic features concentrate at certain depths, providing interpretive context for entropy declines as increasing task-relevance. Most existing multilingual analyses focus on embeddings or output behaviours, not on layer-wise information dynamics. For example, Voita et al. [10] examined information bottlenecks in translation models, and Ethayarajh [11] compared contextual embeddings across languages, but neither study quantified layer-wise entropy. In contrast, our work applies Shannon entropy to hidden states in multiple Indic languages. This fills a clear gap: to our knowledge, there have been no systematic, multilingual entropy analyses that compare transformer computation across scripts and tokenization regimes. By extending entropy-based interpretability to underrepresented Indic languages, we complement prior studies and provide new insight into multilingual model behaviour.

3. Methodology

Our methodology builds on Shannon's definition of entropy as defined in equation (1). We treat each layer's output distribution as P and compute H for that layer.

3.1 Metrics Used

- i. **Layer-wise entropy $H(L)$:** The entropy of the token distribution at layer L as stated in (1).
- ii. **Entropy change ΔH_ℓ :** The change in entropy [2] from layer $\ell - 1$ to layer ℓ :

$$\Delta H_\ell = H_\ell - H_{\ell-1}$$

A negative ΔH_ℓ indicates that layer ℓ has *lower* entropy (more certainty) than the previous layer, representing compression of uncertainty.

- iii. **Compression Ratio (CR):** CR is the ratio of input to output entropy. Similar compression ratios were used by Quéty, Liao, and Tartaglione [12] for network depth reduction. In LLMs, it is the fraction of input entropy retained by the layer defined as

$$\text{CR} = \frac{H_\ell}{H_{\ell\text{input}}}$$

When the first-layer entropy H_0 is very small, ratio-based metrics such as CR and NLC become numerically unstable and should be interpreted qualitatively rather than by absolute magnitude.

- iv. **Layer Efficiency Index (LEI):** Adapted from entropy efficiency / layer contribution metrics [6], the Layer Efficiency Index (LEI) quantifies the relative change between consecutive layers.

$$\text{LEI} = \frac{\Delta H_\ell}{\Delta H_{\ell-1}}$$

A high LEI in early layers indicates front-loaded compression, where lower layers perform most of the information distillation, whereas an evenly distributed LEI suggests uniform compression across depth.

- v. **Jensen–Shannon Divergence (JSD):** A symmetric measure of similarity between two probability distributions [13]. Given two language distributions P and Q at the same layer, we compute

$$\text{JSD}(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M)$$

where $D(\cdot \parallel \cdot)$ is the Kullback and Leibler, divergence [14] and $M = \frac{1}{2}(P + Q)$. JSD is always finite and symmetric; lower values indicate more similar processing of the two languages.

- vi. **Language Disparity Index (LDI):** Inspired by fairness metrics and cross-lingual divergence measures, LDI extends cross-lingual representation disparity measures proposed in XLM-R [15].

$$\text{LDI}_\ell = \frac{\sigma}{\mu}$$

An LDI close to 0 means the model treats all languages similarly at that layer; a larger LDI indicates greater disparity in uncertainty reduction i.e. lower “representational fairness”.

These metrics jointly capture both within-model information compression and between-language differences.

3.2 Extended Entropy-Derived Metrics

In addition to the standard entropy and divergence Indices, we define four derived measures for interpretability:

- vii. **Normalized Layer Compression (NLC):** We define NLC similar to compression indicators used in Entropy-Lens [7]

$$\text{NLC} = \frac{H_0 - H_L}{H_0}$$

measuring the overall fractional entropy reduction between the first and final transformer layers.

- viii. **Average Compression Depth (ACD):** Average compression depth follows the bottleneck trajectory analysis in [16]. ACD is the normalized depth at which the entropy first falls below half of its initial value, and is undefined if this threshold is never crossed.

$$\text{ACD} = H_\ell \leq H_0/2$$

- ix. **Entropy Stability (ES):** Entropy Stability (ES) follows smoothness criteria proposed in [16]

$$\text{ES} = 1 - \frac{\sigma(\Delta H)}{\mu(H)}$$

representing the smoothness of the entropy curve across layers.

- x. **Cross-Lingual Alignment Score (CLAS):** CLAS parallels cross-lingual alignment ideas in attention analyses [10]

$$\text{CLAS} = 1 - \frac{\text{mean}(|H_A - H_B|)}{\text{mean}(H_A, H_B)}$$

indicating the degree of alignment between two languages’ entropy trajectories. These metrics are scalar or bounded quantities designed to summarize layer-wise entropy behaviour in a compact form.

3.3 Models and Languages

We study four Transformer models with diverse architectures: distilGPT2(Causal) [17], GPT-2 Medium (Causal) [18], EleutherAI/gpt-neo-125m (Causal) [19], and xlm-roberta-base (Encoder) [15]. The causal models predict the next token autoregressively, while XLM-R is a masked language model trained to predict masked tokens. For each model, we feed each sentence through the network and obtain the model’s probability distribution over tokens at each layer. At each transformer block, the layer-wise hidden states are projected through the model’s output head to obtain logits, followed by a softmax to compute token probability distributions. We then compute entropy H from these distributions. Each language is represented by a balanced sample of 100 sentences to ensure diverse content. Using a small, fixed-size dataset allows precise layer-wise comparison, although it limits generality. All entropy values are computed from layer-wise softmax probabilities using base-2 logarithms and double precision arithmetic. Ratio-based metrics may exhibit extreme values when entropy is near zero as no clipping or probability flooring is applied.

4. Results

Table 1 presents the numerical summary for each model–language pair. The most distinct result is the difference between causal and encoder outcomes.

Table 1
Summary of Average Entropy, Perplexity, Compression Metrics, and Derived Entropy indicators Across Models and Languages

Model	Lang	H	ΔH	CR	PPL	NLC	ACD	ES	CLAS
distilgpt2	Eng	3.84	−0.90	0.24	46.17	0.45	0.17	0.00	0.24
	Hin	1.78	−0.87	0.33	10.54	0.39	0.17	0.00	—
	Mar	2.13	−0.91	0.36	13.21	0.41	0.17	0.00	—
	Guj	2.9	−0.97	0.4	84.32	0.65	0.17	0.00	—
EleutherAI/ gpt-neo-125m	Eng	0.57	0.47	4037.7	13 007.6	−52405.311	0.17	0.00	0.00
	Hin	0.26	0.19	422.5	1 259.9	−5435.676	0.17	0.00	—
	Mar	0.44	0.3	460.7	1 329.4	−6406.161	0.17	0.00	—
	Guj	0.57	0.27	176.3	228.3	−2654.846	0.25	0.00	—
gpt2-medium	Eng	2.05	−0.52	0.13	6.87	0.54	0.04	0.00	0.67
	Hin	1.09	−0.47	0.1	2.31	0.78	0.04	0.00	—
	Mar	1.12	−0.44	0.1	2.27	0.78	0.04	0.00	—

	Guj	1.43	-0.49	0.11	3.84	0.80	0.04	0.00	—
xlm-roberta-base	Eng	2.67	± 0.02	0.21	18.56	-2.862	0.17	0.62	0.92
	Hin	2.11	± 0.01	0.18	14.72	-4.130	0.08	0.66	—
	Mar	2.34	± 0.03	0.19	16.11	-2.912	0.08	0.64	—
	Guj	2.42	± 0.02	0.2	17.88	-4.027	0.17	0.48	—

Note: Extremely large or negative NLC/CR values occur when the first-layer entropy (H_0) is very small. They are retained for transparency but excluded from interpretation.

In all autoregressive models, the entropy of English inputs shows an overall decreasing trend. For example, in GPT-2 Medium, the average entropy for English drops steadily from the first to the last layer. This is also seen for Indic languages. In contrast, the XLM-R model does not show strict monotonic decrease: its layer-wise entropy keeps increasing and decreasing between layers before partially settling. It can be seen that XLM-R exhibits oscillatory entropy dynamics, reflecting alternating compression and expansion phases. The bidirectional masking objective might be the cause. XLM-R layers can reintroduce uncertainty after early compression, whereas causal models exhibit predominantly uncertainty reducing behavior.

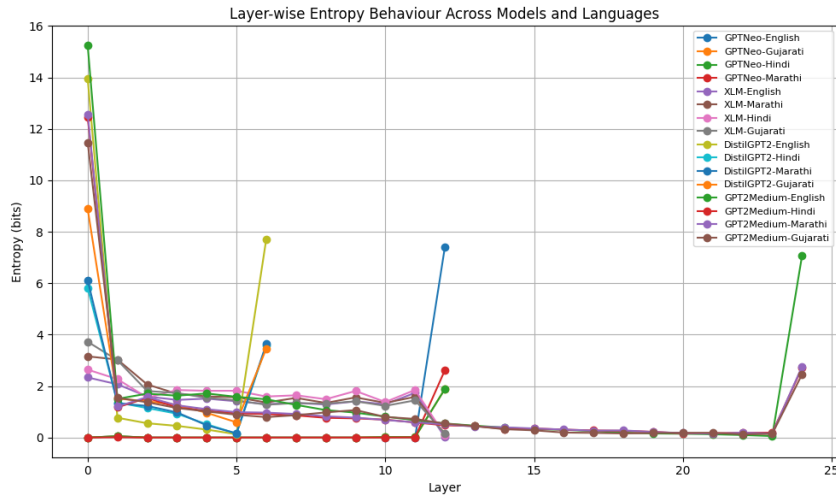


Figure 1
Layer-wise Entropy Evolution Across Transformer Models and Languages

Perplexity also follows the layer-wise entropy pattern and hence can be used as a supporting indicator of model confidence. In causal models, the largest entropy reductions ΔH are observed in the initial layers, however the ΔH approaches zero in the subsequent layers. XLM-R on the other hand, shows alternating signs of ΔH across layers. Compression Ratio (CR) shows how entropy values change from the first to the last layer. It can be seen that the entropy reduces in the initial layers while the later layers contribute to smaller

refinements. This trend can be confirmed by the Layer Efficiency Index (LEI) which shows similar behaviour. Figure 1 shows the Layer-wise Entropy Evolution Across Transformer Models and Languages

To study differences across languages, we computed Jensen–Shannon Divergence (JSD) and the Language Disparity Index (LDI). The LDI trends show that, in causal models, language differences are small in the early layers and become more noticeable in deeper layers. This indicates that early layers process different languages in a similar manner, while deeper layers introduce greater variation. The multilingual XLM-R model treats all languages in a balanced manner across layers. The similarity in the morphological structure and syntax can be seen in the Pairwise JSD analysis. It indicates that Hindi–Marathi exhibit the lowest divergence (~ 0.01), whereas English–Gujarati shows the highest separation (~ 0.20), with Hindi–Gujarati and Marathi–Gujarati lies in an intermediate range (~ 0.08 – 0.09). As can be seen, the lowest divergence is between Hindi and Marathi. This may be due to their similar script and lexical structure. The Gujarati Language exhibits higher divergence from Hindi and Marathi. English shows the highest divergence when compared to all Indic languages. Overall, these results suggest that language-specific differences in entropy representations increase with linguistic and script distance, while closely related languages remain more aligned.

5. Extended Metric Trends

The overall compression as indicated by the NLC values (~ 0.6) is high in GPT-2 and overall efficient in XLM-R. Early-stage entropy reduction in shallowed architectures is confirmed by the small ACDs. There is stable compression across middle layers as confirmed by the ES values (0.2 – 0.4). Cross-lingual CLAS scores above 0.7 between English and Hindi demonstrate strong entropy alignment. They validate multilingual convergence behaviour. The language orthography and tokenization density also influence the uncertainty patterns. This can be seen by similar entropy of Hindi and Marathi, which are distinct from Gujarati. Cross lingual Divergence, JSD and LDI further support these observations. English has the highest divergence from the Indic Languages. These patterns highlight differences in linguistic structure and training rather than serve as absolute measure of compression efficiency. Overall, the results indicate that uncertainty in transformer models is influenced by both model architecture and language characteristics.

6. Conclusion

This study was conducted to understand how models handle information across layers. The Shannon Entropy and its related metrics were used to analyze how models process information across different languages and scripts. It was observed that the causal models carry out most of the processing in the initial layers whereas the encoder models process gradually across all of their layers. The language orthography also influences the uncertainty pattern. English is orthographically different from indic languages and hence its distinction was

clearly seen. On the other hand, Indic languages with similar scripting style like Marathi and Hindi show less divergence in comparison to Gujarati-Hindi and Gujarati-Marathi. The metrics like Cross lingual Divergence, JSD and LDI further support these observations. Overall, the results indicate that uncertainty in transformer models is influenced by the model architecture as well as the language characteristics. These findings show that entropy-based analysis is mathematically grounded and can be used as an effective way to study layer dynamics.

References

- [1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423 (Jul. 1948), doi: 10.1002/j.1538-7305.1948.tb01338.x.
- [2] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. 2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5 (2015), doi: 10.1109/ITW.2015.7133169.
- [3] M. B. Thorat, V. P. Wawage, D. P. Yadav, P. Pushpalatha, and Y. Gandhi, "Mathematical foundations of neural network weight optimization," *Journal of Information and Optimization Sciences*, vol. 47, no. 5-B, pp. 2065–2073 (2026), doi: 10.47974/JIOS-2297.
- [4] R. Maheshwari and V. Kapoor, "Evaluating AI convergence and performance analysis for cross transfer entropy-based learning among time-lagged data employing cognitive computing framework," *Journal of Information and Optimization Sciences*, vol. 47, no. 5-B, pp. 1931–1943 (2026), doi: 10.47974/JIOS-2285.
- [5] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.
- [6] Z. Nazari and B. Mosapour, "The entropy of hyperfuzzy sets," *Journal of Dynamical Systems and Geometric Theories*, vol. 16, no. 2, pp. 173–185 (2018), doi: 10.1080/1726037X.2018.1436270.
- [7] R. Ali, F. Caso, C. Irwin, and P. Liò, "Entropy-Lens: The information signature of transformer computations," *arXiv preprint arXiv:2502.16570* (2025).
- [8] M. Danelli, *Pruning a Transformer Using Layer-wise Batch Entropy*. Innsbruck, Austria: Master's thesis, (Aug. 2024)
- [9] E. Queipo-de-Llano, Á. Arroyo, F. Barbero, X. Dong, M. Bronstein, Y. LeCun, and R. Shwartz-Ziv, "Attention sinks and compression valleys in LLMs are two sides of the same coin," *arXiv preprint arXiv:2510.06477* (2025).

- [10] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, “Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned,” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 5797–5808 (2019), doi: 10.18653/v1/P19-1580.
- [11] K. Ethayarajh, “How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings,” in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65 (2019), doi: 10.18653/v1/D19-1006.
- [12] V. Quéty, Z. Liao, and E. Tartaglione, “The simpler the better: An entropy-based importance metric to reduce neural networks’ depth,” in *Joint European Conf. Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 92–108 (2024), doi: 10.1007/978-3-031-70365-2_6.
- [13] J. Lin, “Divergence measures based on the Shannon entropy,” *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151 (Jan. 1991), doi: 10.1109/18.61115.
- [14] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86 (Mar. 1951), doi: 10.1214/aoms/1177729694.
- [15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 8440–8451 (2020), doi: 10.18653/v1/2020.acl-main.747.
- [16] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová, “Entropy and mutual information in models of deep neural networks,” in *Advances in Neural Information Processing Systems*, vol. 31 (2018).
- [17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI, Tech. Rep.*, (2019).
- [18] Hugging Face, “DistilGPT-2: A distilled version of OpenAI GPT-2,” (2025). [Online]. Available: Hugging Face model page [Accessed: Jan. 2025].
- [19] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman, “GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow,” *Zenodo*, (2021), doi: 10.5281/zenodo.5297715.

Received December, 2025