IQ-NET for Sequences: A Holistic and Interpretable Evaluation Framework

Yasser Sajjadi [1]

[1] *Independent Researcher, yassersajjadi@gmail.com*

Abstract: Standard benchmarks for deep learning models across various domains are often confined to a narrow set of metrics like accuracy, which fail to capture a model's full character. To address this, we introduce the **IQ-NET** project, a suite of evaluation frameworks designed to provide a multi-faceted assessment of deep learning architectures tailored for different data modalities. This paper presents the first installment of this project: **IQ-NET for Sequences**. This framework is specifically designed to holistically evaluate sequence modeling architectures using nine distinct metrics: Processing IQ, Stability, Adversarial Robustness Index (ARI), Learning Speed Index (LSI), Efficiency Scalability Index (ESI), Efficiency Complexity Index (ECI), Generalization Transfer Index (GTI), Output Novelty Ratio (ONR), and Reasoning Index (REI). Each metric probes a specific aspect of model performance, from its internal representation quality to its robustness and efficiency. We apply IQ-NET for Sequences to benchmark seven architectures—CNN, TCN, LSTM, GRU, Transformer, Longformer, and the Zarvan model—on the IMDB sentiment analysis task. Our results demonstrate that this framework provides a much richer perspective than standard benchmarks. For instance, TCN achieves the highest final IQ Score due to its balanced profile, while the Transformer exhibits superior internal representation quality. This paper details the theoretical foundations of each metric, the experimental setup, and a thorough analysis, showcasing how IQ-NET for Sequences can guide researchers in understanding and improving sequence models. The code and experimental setups are available at https://github.com/systbs/iq-net/.

Keywords: Deep Learning, Evaluation Framework, Benchmarking, Sequence Modeling, Natural Language Processing (NLP), Model Interpretability, Computational Efficiency, Adversarial Robustness, Representation Quality, IQ-NET, Transformer, TCN.

---

1. Introduction

The evaluation of deep learning models, whether in Natural Language Processing, Computer Vision, or signal processing, has traditionally been dominated by performance-centric metrics like Accuracy and F1-Score. This narrow focus treats the model as a black box, revealing *what* it accomplishes but providing little insight into *how* or *why*. Key characteristics essential for practical application—such as computational cost, parameter efficiency, and robustness—are often overlooked. This gap in evaluation methodology hinders the development of truly robust and efficient AI systems.

To create a more comprehensive and standardized evaluation paradigm, we are developing the **IQ-NET** project. The goal of this project is to create a suite of holistic and interpretable frameworks, with each framework tailored to the unique challenges of a specific data modality. This paper introduces the first official module of this initiative: **IQ-NET for Sequences**.

IQ-NET for Sequences is a framework specifically designed for the nuanced evaluation of sequence modeling architectures. It moves beyond simple performance metrics to quantify a model's effectiveness and efficiency through nine carefully designed indices, each grounded in established scientific principles.

The core contributions of this paper are:

1. The formal introduction of the **IQ-NET for Sequences** framework and its nine constituent metrics.
2. A comprehensive benchmark of seven prominent sequence modeling architectures: CNN, TCN, LSTM, GRU, Transformer, Longformer, and Zarvan.
3. A detailed analysis demonstrating how our framework provides deeper insights into model behavior, revealing trade-offs invisible to standard protocols and offering clear directions for future research.

2. The IQ-NET for Sequences Framework

The IQ-NET for Sequences framework consists of nine metrics, each targeting a specific and interpretable aspect of model performance.

2.1. Processing IQ (Proc. IQ)

- **Definition**: Processing IQ measures a model's ability to transform input representations into meaningful, task-relevant outputs. It is a composite score evaluating the quality of the entire representational pipeline, from initial embeddings to the final hidden states and output logits.
- **Formula**: The metric is computed as the product of three components: Input-to-Hidden Centered Kernel Alignment (ITC), Hidden-to-Output CKA (CCI), and the Ratio of Explained Variance from PCA (RP).

$$\text{Proc. IQ} = \text{ITC} \cdot \text{CCI} \cdot \text{RP}$$

where:

- \`$\text{ITC} = \text{CKA}(h_{\text{initial}}, h_{\text{final}})$\`
- \`$\text{CCI} = \text{CKA}(h_{\text{final}}, y)$\`
- \`$\text{RP} = \frac{\text{EVR\_final}}{\text{EVR\_initial}+\text{EVR}_{\text{final}}+\epsilon}$\`

- **Scientific Rationale**: This metric is defensible because its components are based on robust and widely accepted techniques.

- **CKA (Centered Kernel Alignment)** is a powerful method for comparing representation spaces in neural networks, as it is invariant to orthogonal transformations and can capture non-linear similarities. It effectively assesses how well a model transforms inputs into structured hidden states and then into outputs. The use of CKA is supported by work from Kornblith et al. (2019).
- **PCA (Principal Component Analysis)** is a standard technique for dimensionality reduction and evaluating information density in high-dimensional data. The RP term quantifies the preservation or concentration of task-relevant variance in the final hidden states relative to the initial ones, reflecting the model's ability to encode useful information.

- **Insight for Researchers**: A high Proc. IQ suggests that the model has learned a well-structured internal pipeline where representations are progressively and effectively refined. A low score, even with high accuracy, might indicate that the model's internal representations are messy or that it relies on brittle shortcuts. This metric guides researchers to focus on architectures that promote better representational quality, potentially leading to improved generalization and robustness.

2.2. Stability

- **Definition**: Stability quantifies the consistency of a model's processing quality (its Proc. IQ score) across different levels of input difficulty. It measures how reliably the model performs its internal transformations.
- **Formula**: It is calculated as the inverse of the coefficient of variation of Proc. IQ scores computed over several difficulty buckets.

$$\text{Stability} = \max\left(0{,}1.0 - \frac{\text{std(IQ\_scores)}}{\text{mean(IQ\_scores)} + \epsilon}\right)$$

where \`$IQ_{scores}$\` are the Proc. IQ values for different data subsets.

- **Scientific Rationale**: The use of the **coefficient of variation (textstd/textmean)** is a standard statistical method for measuring relative variability, making it scale-invariant and suitable for comparing models with different performance ranges. By subtracting this value from 1, we create an intuitive score where higher values denote higher stability.
- **Insight for Researchers**: Stability is a proxy for reliability. A model with high stability can be trusted to perform consistently across diverse real-world inputs, whereas a model with low stability may be unpredictable. This metric encourages researchers to design models that are not just accurate on average but are also dependable.

2.3. Adversarial Robustness Index (ARI)

- **Definition**: ARI evaluates a model's resilience to perturbations in its input data. It measures how much the model's loss increases when random noise is introduced into the input sequences at levels of 5%, 10%, and 20%.

- **Formula**: ARI is the sum of the ratios of base loss to adversarial loss, tested across multiple noise levels.

$$\text{ARI} = \sum_{\text{noise\_level}\in\{0.05,0.1,0.2\}} \frac{Loss_{\text{base}}}{Loss_{\text{adversarial}} + \epsilon}$$

- **Scientific Rationale**: The principle that small input perturbations should not drastically alter a model's output is a cornerstone of robust machine learning, as established in seminal works like Szegedy et al. (2014) and Madry et al. (2018). Testing at multiple noise levels ensures a comprehensive assessment of robustness across a range of severities. The ratio-based formula directly quantifies performance degradation under attack.

- **Insight for Researchers**: ARI provides a direct measure of a model's reliability in noisy, real-world environments. A low ARI score is a significant red flag, indicating that the model may fail unexpectedly when faced with imperfect data. It guides research towards regularization techniques, data augmentation, or architectural choices that enhance robustness.

2.4. Learning Speed Index (LSI)

- **Definition**: LSI measures the rate of error reduction during training, quantifying how quickly and efficiently a model learns.

- **Formula**: It is calculated as the sum of normalized error reductions over epochs, where earlier improvements are weighted more heavily.

$$\text{LSI} = \sum_{t=1}^{T} \frac{L_0 - L_t}{t + 1}$$

where L_t is the training loss at epoch t and T is the total number of epochs.

- **Scientific Rationale**: The formula captures the cumulative learning progress, normalizing by the epoch index (t+1) to account for the natural diminishing returns in training. This approach reflects realistic learning dynamics and is particularly useful for comparing models with different convergence behaviors, a concept discussed in works on training efficiency like Bottou (2012).

- **Insight for Researchers**: LSI is a critical metric for evaluating training efficiency. A model with a high LSI converges faster, saving significant computational resources and time. This is especially important for large-scale NLP tasks. It can help identify more efficient architectures or hyperparameter settings.

2.5. Efficiency Scalability Index (ESI)

- **Definition**: ESI assesses a model's computational scalability, specifically how its inference time is affected by increasing input sequence length.

- **Formula**: ESI is defined as the inverse of the product of the inference time for a medium-length sequence and the ratio of logarithmic slopes of inference times.

$$\text{ESI} = \frac{1}{r \cdot t_2 + \epsilon}$$

where r is the ratio of slopes $m_{3/2}$ to $m_{1/2}$, and t_2 is the inference time for sequence length 512.

- **Scientific Rationale**: Using logarithmic slopes is a standard method for analyzing computational complexity, especially for sequence models where complexity can be non-linear (e.g., quadratic in Transformers). The formula is designed to reward models with near-linear scaling (a slope ratio r close to 1) and low absolute inference times, which are key goals in efficient model design.

- **Insight for Researchers**: ESI is crucial for selecting models for applications that handle long documents or require real-time processing. It provides a clear, quantitative measure of how a model's performance will scale, allowing a researcher to anticipate bottlenecks and choose architectures (like Longformer or Zarvan) that are explicitly designed for efficiency on long sequences.

2.6. Efficiency Complexity Index (ECI)

- **Definition**: ECI measures a model's parameter efficiency by calculating its classification accuracy per million trainable parameters.

- **Formula**:

$$\text{ECI} = \frac{\text{Accuracy}}{\text{Number of Parameters}/10^6}$$

- **Scientific Rationale**: Parameter efficiency is a key consideration in deep learning, especially for deployment on resource-constrained devices. This metric directly relates model performance to its complexity (size), rewarding compact models that achieve high accuracy. This principle is widely discussed in foundational deep learning literature.

- **Insight for Researchers**: ECI helps researchers evaluate the trade-off between model size and performance. It encourages the development of smaller, more efficient models without sacrificing accuracy, which is a major research direction in fields like model pruning, quantization, and knowledge distillation.

2.7. Generalization Transfer Index (GTI)

- **Definition**: GTI evaluates a model's ability to generalize from the training data to unseen test data by comparing their respective losses.

- **Formula**: It is the ratio of the average training loss to the average test loss.

$$\mathrm{GTI} = \frac{\text{Train Loss}}{\text{Test Loss} + \epsilon}$$

- **Scientific Rationale**: The ratio of training to test error is a standard and fundamental metric for diagnosing overfitting in machine learning. A ratio close to 1 indicates good generalization, while a very low ratio suggests the model has overfit the training data. The use of average losses across batches ensures a robust estimation. This concept is central to texts like Goodfellow et al. (2016).
- **Insight for Researchers**: GTI provides a clear signal of a model's generalization capability. A low GTI is a warning sign of overfitting, prompting researchers to apply stronger regularization, gather more data, or simplify the model architecture.

2.8. Output Novelty Ratio (ONR)

- **Definition**: ONR measures the diversity and magnitude of a model's outputs (logits).
- **Formula**: It is the product of the entropy of the output distribution and the mean absolute output value.

$$\mathrm{ONR} = \mathrm{Entropy}(p_y) \cdot \mathrm{Mean}(|y|)$$

where p_y is the normalized histogram of model outputs.

- **Scientific Rationale**: **Entropy**, a cornerstone of information theory, is used here to measure the diversity of the model's predictions. High entropy indicates the model is using a wider range of its output space, which can be a sign of a more robust classifier. Multiplying by the **mean absolute output** incorporates prediction confidence, as higher magnitude logits often correspond to more confident predictions.
- **Insight for Researchers**: ONR provides a view into the "expressiveness" of the model's output layer. A model with very low ONR might be producing collapsed or low-confidence predictions, even if they are often correct. This can guide researchers to investigate issues like activation saturation or poor initialization in the final layers.

2.9. Reasoning Index (REI)

- **Definition**: REI evaluates a model's nascent ability to refine its own predictions in an iterative manner. It measures the performance improvement gained by feeding a model's own projected conclusions back into its input embeddings for a second pass.
- **Formula**:

$$\mathrm{REI} = \max\left(0, \frac{\text{Loss\_single} - \text{Loss\_multi}}{\text{Loss\_single} + \text{Loss\_multi} + \epsilon}\right)$$

- **Scientific Rationale**: The concept of iterative refinement is inspired by studies in meta-learning and biologically plausible deep learning, where systems refine their understanding over time. The normalized difference formula ensures the metric is scale-invariant and captures relative improvement, making it comparable across different models.

- **Insight for Researchers**: REI is a forward-looking metric that probes a model's capacity for self-correction and reasoning—a key feature for advanced AI. While current models may score low, it provides a valuable benchmark to measure progress on more complex reasoning tasks and encourages the development of architectures with feedback loops or other iterative mechanisms.

3. Final Score Calculation

To provide a single, interpretable summary of a model's overall quality, the individual metrics are aggregated into a **Final Score** using a weighted average.

The weights are assigned based on the perceived importance of each metric for general-purpose NLP model evaluation. Core aspects like representation quality (Proc. IQ), robustness (ARI), and learning speed (LSI) are given high importance. Secondary aspects like parameter efficiency (ECI) and generalization (GTI) receive medium importance, while supplementary aspects like stability and scalability receive lower importance.

Table 1: Weights assigned to each metric in the IQ-NET for Sequences framework. The weights are scaled to balance the raw metric values and reflect their designated importance.

Metric	Weight	Importance
Proc. IQ	359.0	High
Stability	35.0	Low
ARI	34.0	High
LSI	187.0	High
ESI	0.043	Low
ECI	174.0	Medium
GTI	50.0	Medium
ONR	11.0	Medium
REI	126.0	Low

The Final Score is computed using the following formula:

$$\text{Final Score} = \left(\frac{\sum_{k \in \text{Metrics}} w_k \cdot v_k}{\sum_{k \in \text{Metrics}} w_k} \right) \cdot 160$$

The division by the sum of weights normalizes the score, making it scale-invariant and comparable across models. The final scaling factor of 160 is chosen empirically to map the score to an intuitive range, typically between 0 and 100.

---

4. Experimental Setup

4.1. Dataset and Preprocessing

The experiments were conducted on the **IMDB dataset**, a standard benchmark for binary sentiment classification, loaded using the datasets library. For metrics requiring varied difficulty (Stability, Proc. IQ), the training data was sorted by sequence length and partitioned into 10 "difficulty buckets". A vocabulary of 20,000 tokens was constructed from the training texts.

4.2. Models and Hyperparameters

Seven sequence modeling architectures were evaluated:

- **Baselines:** CNN, TCN, LSTM, GRU.
- **Transformer-based:** Transformer, Longformer.
- **Novel Architecture:** Zarvan.

All models were trained and evaluated using a consistent set of hyperparameters to ensure fair comparison. The key parameters are detailed in Table 2. The Adam optimizer with a learning rate of 0.001 was used for all models. All experiments were run on a CUDA-enabled GPU device.

Table 2: Key hyperparameters used for all model configurations during training and evaluation.

Hyperparameter	Value
Vocabulary Size	20000
Embedding Dim	128
Hidden Dim (RNNs)	256
Num Layers	2
Num Heads (Attn)	4
Num Classes	2
Num Filters (CNN)	100
Batch Size	16
Max Epochs	5
Max Sequence Len	4096

4.3. Reproducibility

To ensure the reproducibility of our results, a fixed random seed (seed=42) was used for all random processes in PyTorch, NumPy, and Python's random module. The use of standardized datasets and clearly defined mathematical formulations further contributes to the framework's reproducibility.

5. Results and Analysis

The comprehensive evaluation of the seven models using the IQ-NET for Sequences framework yielded the results summarized in Table 3.

Table 3: The final IQ-NET for Sequences report card. Models are ranked by their final aggregated score. The table shows the raw score for each of the nine metrics alongside the final score.

Model	**Final Score**	**Proc. IQ**	**Stability**	**ARI**	**LSI**	**ESI**	**ECI**	**GTI**	**ONR**	**REI**
TCN	83.6630	0.1484	0.9472	2.8966	0.4841	95.4485	0.3802	1.3297	5.9158	0.2628
Transformer	81.2227	0.2755	0.9033	2.9356	0.4430	164.4344	0.3451	0.8513	5.6277	0.0852
Zarvan	78.5745	0.2291	0.6838	2.9118	0.5286	230.5158	0.3098	0.7688	5.0911	0.1352
CNN	69.9885	0.1671	0.8527	2.7733	0.3328	783.2341	0.3684	0.9409	3.0522	0.0174
GRU	61.5644	0.1658	0.6230	2.9233	0.2952	156.2350	0.2988	1.1766	0.7992	0.1055
Longformer	60.5903	0.2153	0.8348	3.0005	0.3025	61.2647	0.1177	0.9961	2.1593	0.0624
LSTM	56.6507	0.1861	0.7182	2.8948	0.2045	130.2392	0.2529	1.2022	0.5106	0.0131

The results reveal a nuanced story. **TCN** emerges as the top-performing model overall, not by dominating any single metric, but by demonstrating a strong, balanced profile across the board. It scores well on LSI, ECI, GTI, and ONR. In contrast, the **Transformer** model, which ranks second, owes its high score to its exceptional performance on Proc. IQ (0.2755), the highest of any model, indicating superior representational quality. The novel **Zarvan** architecture achieves the highest LSI score (0.5286), confirming its efficient learning dynamics.

The radar charts in Figure 1 and Figure 2 visualize these complex trade-offs, providing an intuitive understanding of each model's "personality."

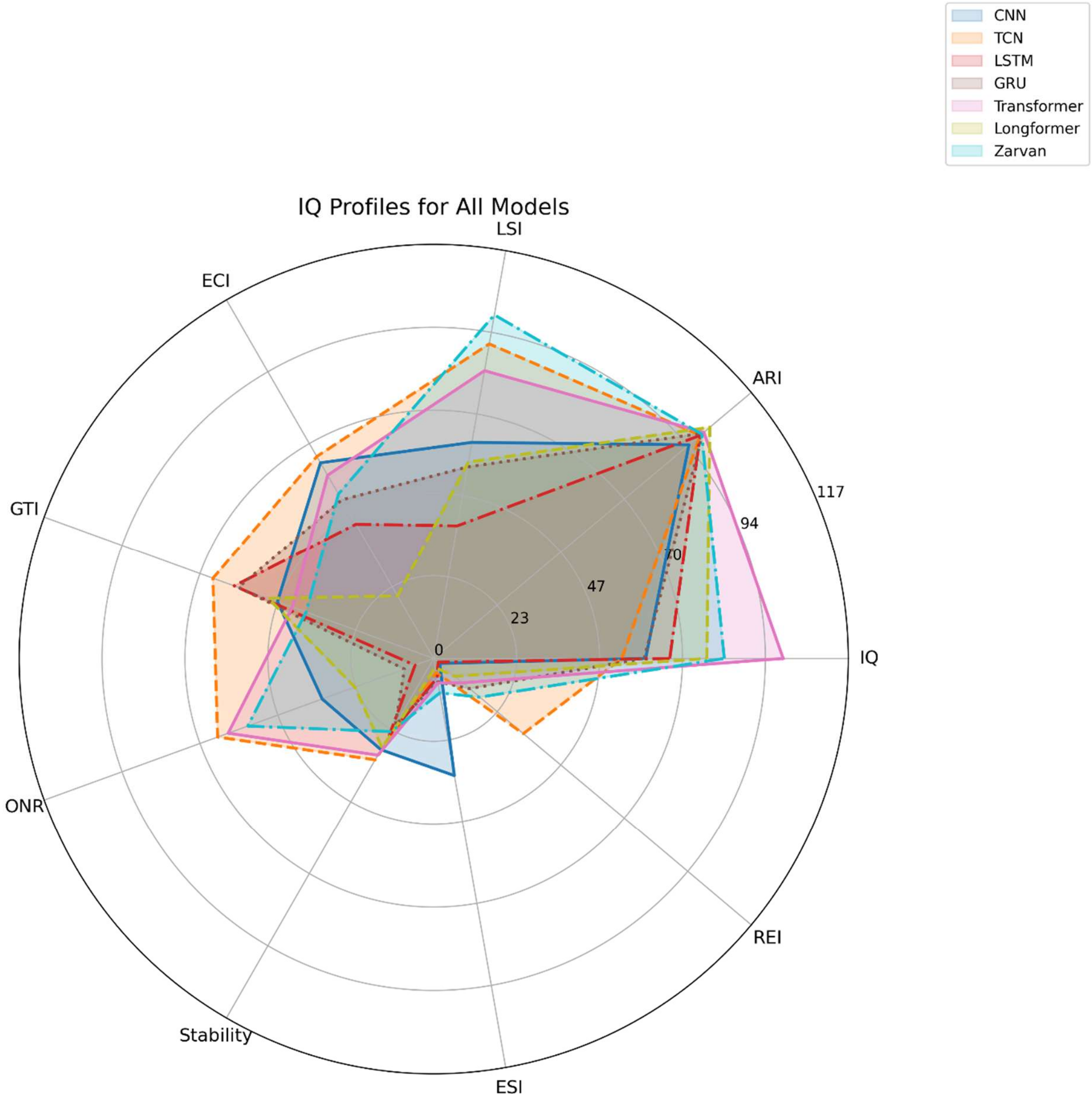

Figure 1: The combined IQ Profiles for all seven models. This chart visually summarizes the relative strengths and weaknesses of each architecture, highlighting the trade-offs between different performance aspects. For example, the trade-off between Transformer's high Proc. IQ (pink line) and TCN's balanced profile (orange line) is immediately apparent.

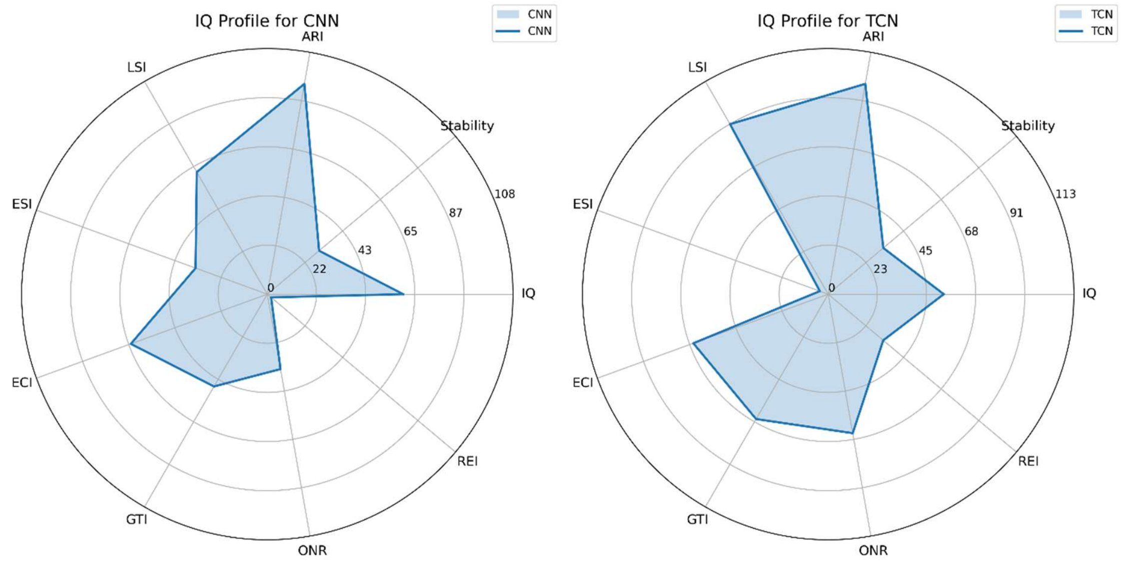

Figure 2: Individual IQ Profiles for each model. These charts reveal the unique performance signature of each architecture. For example, the profile for CNN clearly shows its strengths lie in efficiency and scalability (ESI, ECI), while Longformer's profile highlights its high robustness (ARI).

6. Discussion

The primary value of the IQ-NET for Sequences framework lies in the depth of insight it offers compared to standard benchmarks. A standard accuracy-based evaluation would rank TCN (86.9%) and CNN (86.3%) as the top models, with Transformer (83.0%) and Zarvan (81.3%) trailing. While not incorrect, this view is incomplete.

IQ-NET provides a multi-dimensional analysis that explains *why* these models perform as they do. For instance:

- **CNN's Efficiency**: Although its accuracy is high, IQ-NET reveals its true strength: exceptional scalability (ESI score of 783) and parameter efficiency (ECI score of 0.368). This makes it a prime candidate for applications where computational resources are a major constraint.
- **Transformer's Quality**: Despite lower accuracy than TCN, its top score in Proc. IQ (0.2755) suggests its architecture is fundamentally better at learning high-quality, structured representations from data. This may translate to better performance on more complex tasks.
- **Longformer's Robustness**: Longformer achieves the highest ARI score (3.0005), making it the most robust model against input perturbations. This is a critical feature for real-world systems that must handle noisy data, a fact completely missed by standard evaluation.

- **Zarvan's Learning Speed**: The new Zarvan architecture demonstrates the fastest learning rate (LSI of 0.5286). This indicates a highly efficient architecture in terms of training convergence, providing a strong foundation for future development.

This framework allows a researcher to move beyond a simple leaderboard and make decisions based on a model's character. For a mobile application, CNN's high ECI and ESI might be decisive. For a mission-critical system where reliability is paramount, Longformer's high ARI would be a key selling point. For pure research into representation learning, the Transformer's high Proc. IQ makes it the most interesting subject for further study.

7. Conclusion

In this paper, we introduced **IQ-NET for Sequences**, the first module in a broader initiative to create holistic and interpretable evaluation frameworks for deep learning. By moving beyond traditional accuracy-based metrics, IQ-NET for Sequences provides a multi-faceted view of a model's performance, encompassing its internal processing quality, efficiency, scalability, robustness, and learning dynamics. Our comprehensive benchmarking of seven models demonstrates that the framework can successfully uncover nuanced trade-offs and provide deep insights into the unique character of each architecture.

As the first installment of the IQ-NET project, this work lays the foundation for a more comprehensive approach to model evaluation. Future work will focus on expanding the project by developing and releasing specialized frameworks for other data types, including **IQ-NET for Images**, **IQ-NET for Audio**, **IQ-NET for Signals**, and **IQ-NET for Video**.

8. References

[1] Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). *Similarity of Neural Network Representations Revisited*. International Conference on Machine Learning (ICML).

[2] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

[3] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[4] Everitt, B. S., & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. Cambridge University Press.

[5] Szegedy, C., Zaremba, W., Sutskever, I., et al. (2014). *Intriguing Properties of Neural Networks*. International Conference on Learning Representations (ICLR).

[6] Madry, A., Makelov, A., Schmidt, L., et al. (2018). *Towards Deep Learning Models Resistant to Adversarial Attacks*. ICLR.

[7] Bottou, L. (2012). *Stochastic Gradient Descent Tricks*. Neural Networks: Tricks of the Trade.

[8] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention is All You Need*. Advances in Neural Information Processing Systems (NeurIPS).

[9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep Learning*. Nature.

[10] Shannon, C. E. (1948). *A Mathematical Theory of Communication*. Bell System Technical Journal.

[11] Bengio, Y., Lee, H., & Larochelle, H. (2015). *Towards Biologically Plausible Deep Learning*. arXiv preprint arXiv:1502.04156.

[12] Saaty, T. L. (2008). *Decision Making with the Analytic Hierarchy Process*. International Journal of Services Sciences.

[13] Sajjadi, Y. (2025). *Zarvan: An Efficient Gated Architecture for Sequence Modeling with Linear Complexity*. Preprints.org. doi:10.20944/preprints202507.2512.v1.