Oleksandr Paladiiev*

ENCODER-DECODER ARCHITECTURES AND THEIR UNIVERSALITY

The article provides a systematic analysis of encoder–decoder architectures as a foundational approach to conditional generation tasks of the form "input X –> output Y." It refines the concept of the universality of these architectures in three interrelated dimensions: as the universality of the architectural contract between representation and generation, as the universality of approximation capacity for a broad class of sequence-to-sequence mappings, and as practical universality constrained by computational cost, scaling, and context length. The study traces the evolution of this approach from RNN-based seq2seq models with a fixed vector bottleneck to attention mechanisms, Transformer encoder–decoder architectures, and pretrained seq2seq models such as T5, BART, and PEGASUS. Special attention is devoted to the formalization of the conditional factorization p(y|x), the role of self-attention, cross-attention, and positional encodings, as well as the computational limitations of full attention. The article also examines modern extensions of the paradigm, including long-context models, retrieval-augmented generation, and multimodal systems. Based on a comparison of encoder–decoder, encoder-only, decoder-only, and retrieval-augmented approaches, practical recommendations are formulated for architecture selection depending on the task type, input structure, and resource constraints.

Keywords: *encoder–decoder, sequence-to-sequence, attention, cross-attention, transformer, universal approximation, long context, retrieval-augmented generation, multimodal models.*

Tabl. 2. Fig. 3. Form. 4. Lit. 20.

DOI: 10.32752/1993-6788-2026-1-296-413-424

Peer-reviewed, approved and placed: *19.02.2026*

**** https://orcid.org/0000-0003-4475-1220***

Олександр О. Паладієв

АРХІТЕКТУРИ ЕНКОДЕР–ДЕКОДЕР ТА ЇХ УНІВЕРСАЛЬНІСТЬ

У статті здійснено системний аналіз encoder-decoder архітектур як базового підходу до задач умовної генерації типу «вхід X –> вихід Y». Уточнено зміст поняття універсальності цих архітектур у трьох взаємопов'язаних вимірах: як універсальності архітектурного контракту між репрезентацією та генерацією, як універсальності апроксимаційної здатності для широкого класу sequence-to-sequence відображень і як практичної універсальності, обмеженої обчислювальною вартістю, масштабуванням і довжиною контексту. Простежено еволюцію підходу від RNN-based seq2seq моделей із фіксованим векторним bottleneck до attention-механізмів, Transformer encoder-decoder архітектур і попередньо натренованих seq2seq моделей типу T5, BART, PEGASUS. Окрему увагу приділено формалізації умовного розкладу p(y|x), ролі self-attention, cross-attention і positional encodings, а також обчислювальним обмеженням повної уваги. Розглянуто сучасні розширення парадигми, зокрема long-context моделі, retrieval-augmented generation та мультимодальні системи. На основі порівняння encoder–decoder, encoder-only, decoder-only і retrieval-augmented підходів сформульовано практичні рекомендації щодо вибору архітектури залежно від типу задачі, структури вхідних даних і ресурсних обмежень.

Ключові слова: енкодер-декодер, sequence-to-sequence, attention, cross-attention, transformer, універсальне наближення, довгий контекст, retrieval-augmented generation, мультимодальні моделі.

* International European University, Kyiv, Ukraine.

Problem Statement. In conditional generation tasks of the form "input X –> output Y" (machine translation, abstractive summarization, speech recognition, image captioning, code generation from descriptions, etc.), two distinct subproblems naturally arise: (1) constructing a representation of the input signal and (2) generating an output sequence conditioned on this representation. Early sequence-to-sequence formulations cast this as two networks that optimize the conditional likelihood of the output given the input, while "universality" was interpreted as the ability of a single topology to solve a broad class of transformations between variable-length sequences under minimal assumptions about their structure (Cho K. et al., 2014) [1].

From this perspective, the encoder–decoder paradigm constitutes a universal architectural contract, but not necessarily a universally optimal infrastructural choice in every production setting. In recent practice, there has been a shift toward decoder-only models in many generative applications, driven by the simplicity of the prompt –> completion interface and the scalability of causal language modeling. However, studies that compare encoder–decoder and decoder-only architectures in large-scale regimes indicate that encoder–decoder models can remain competitive – and even advantageous–under certain conditions, particularly in terms of inference efficiency and conditional tasks (Brown T. B., 2020) [2].

Accordingly, it is appropriate to distinguish at least two senses of "universality":

1. Universality of expressivity (the theoretical approximation capacity of a family of models to realize broad classes of sequence-to-sequence mappings);

2. Universality of application (the practical dominance of a given topology as an optimal trade-off among accuracy, training/inference cost, context length, and requirements for knowledge updateability) (Yun C. et al., 2019) [3].

The objective of this work is to provide a systematic account of encoder–decoder architectures as a paradigm for conditional generation and to critically analyze their "universality" in both theoretical and practical senses, with a focus on Transformer-based models and their modern extensions.

The research objectives include:

1. To refine the notion of "universality" for encoder–decoder architectures (contract/ expressivity/ practice) and to demonstrate why these interpretations are not equivalent (Yun C. et al., 2019) [3];

2. To trace the evolution from RNN-based sequence-to-sequence models with a fixed bottleneck to attention-based models and Transformer encoder–decoder architectures, as well as to pretrained seq2seq (text-to-text) approaches (Cho K. et al., 2014) [1];

3. To formalize the factorization $p(y \mid x)$, describe self-attention and cross-attention mechanisms and the role of positional encodings, and to discuss computational complexity as a key practical constraint (Vaswani A. et al., 2017) [4];

4. To compare encoder–decoder architectures with alternative topologies (encoder-only, decoder-only, retrieval-augmented) and to formulate practical recommendations for architecture selection given task classes and resource constraints (Devlin J. et al., 2018) [5].

Analysis of Recent Research and Publications. The issue of the development and universality of encoder-decoder architectures is actively being developed in the world scientific discourse as a fundamental pattern for conditional sequence generation. A

significant contribution to the formation of this paradigm was made in the works of K. Cho (Cho K. et al., 2014) [1] and I. Sutskever (Sutskever I. et al., 2014) [6], who laid the foundations for training models to optimize the conditional probability of output under minimal assumptions about the structure of sequences. An important step in overcoming the structural limitations of a fixed vector ("bottleneck") was the research of D. Bahdanau (Bahdanau D. et al., 2014) [7] and T. Luong (Luong T. et al., 2015) [13], who introduced the attention mechanism as a dynamic interface for interaction between the encoder and decoder. A revolutionary transition to parallelized computing was the work of A. Vaswani et al. (Vaswani A. et al., 2017) [4], who presented the Transformer architecture, which unified the roles of the encoder and decoder through self-attention and cross-attention mechanisms. Further development of the topic in the works of S. Raffel (Raffel C. et al., 2019) [8] and M. Lewis (Lewis M. et al., 2019) [17] demonstrated the effectiveness of the unified "text –> text" approach (T5 and BART models), which confirmed the practical versatility of these architectures for a wide range of tasks. Modern research by P. Lewis Bahdanau D. et al., 2014) [7] and I. Beltagy (Beltagy I. et al., 2020) [9] focuses on scaling architectures for working with long context and integrating external memory through retrieval generation (RAG), which expands the boundaries of the encoder–decoder paradigm in complex information environments.

The purpose of the article. The objective of this article is to provide a systematic account of encoder–decoder architectures as a paradigm for conditional generation and to critically analyze their "universality" in both theoretical (approximation capacity) and practical (efficiency and scalability) senses, with a primary focus on Transformer-based models and their modern extensions.

Presentation of the Main Research Material. The first "canonical" sequence-to-sequence systems were built on a simple principle: the encoder compresses the input into a fixed-length vector, and the decoder expands it into the output. This scheme was conceptually clean and general, but quickly revealed a structural limitation: the fixed vector becomes a bottleneck, especially for longer and more complex inputs. This issue was explicitly identified in works that introduced the attention mechanism for neural machine translation (Sutskever I. et al., 2014) [6].

The attention mechanism fundamentally transformed the interface between encoder and decoder: instead of a single "code," the decoder gained access to the entire sequence of encoder hidden states and learned to select relevant fragments at each step of generation, which was interpreted as a form of soft alignment between input and output tokens. Subsequent work systematized variants of attention (global, local, etc.), establishing it as a core architectural primitive Bahdanau (Bahdanau D. et al., 2014) [7].

The Transformer marked the next major transition: it eliminated recurrence and convolutions as the primary mechanisms for sequence modeling, replacing them with multi-head self-attention and feed-forward blocks. At the same time, the encoder–decoder topology was preserved (masked self-attention in the decoder combined with cross-attention to encoder outputs), but became significantly more amenable to parallelization during training (Vaswani A. et al., 2017) [4].

A separate line of evolution involved the unification of tasks through encoder–decoder pretraining. The text-to-text paradigm (T5) demonstrated that

many NLP tasks can be reformulated within a unified "text –> text" framework, where a bidirectional encoder supports "understanding," and an autoregressive decoder handles generation. Denoising sequence-to-sequence approaches (BART) and task-specific objectives for summarization (PEGASUS) further reinforced this paradigm Raffel (Raffel C. et al., 2019) [8].

Subsequently, "universality" became increasingly dependent on two key engineering extensions.

First, long-context modeling: standard self-attention exhibits quadratic complexity with respect to sequence length, motivating the development of sparsity patterns (e.g., Longformer) and generative variants such as LED for long-document sequence-to-sequence tasks (Beltagy I. et al., 2020) [9].

Second, retrieval as an "external memory encoder": retrieval-augmented generation (RAG) and approaches such as Fusion-in-Decoder (FiD) showed that, for knowledge-intensive tasks, it is effective to separate knowledge access (via indexing and retrieval) from generation (the seq2seq core), with cross-attention in the decoder serving as the mechanism for aggregating multiple information sources (Lewis P. et al.. 2020) [10].

Finally, the encoder–decoder paradigm naturally extended to multimodal settings: vision –> text (image captioning), audio –> text (end-to-end ASR), as well as modern vision–language models (VLMs), where the "bridge" between modalities is implemented via controlled cross-attention or compact querying/bridging modules (Vinyals O. et al., 2014) [11].

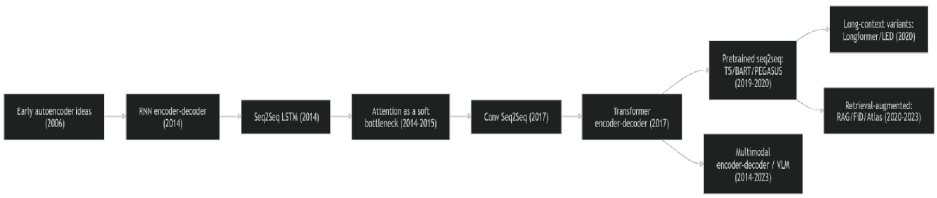

Fig. 1. **Schematic evolution of encoder–decoder approaches (architectural "inflection points")**

Formalization and Variations of the Architecture

Basic factorization. Let the input be a sequence $x = (x_1, ..., x_{T_x})$, and the output $y = (y_1, ..., y_{T_y})$. A typical sequence-to-sequence model defines the conditional distribution via an autoregressive factorization:

$$p(y \mid x) = \prod_{t=1}^{T_y} p\,(y_t \mid y_{<t}, x). \tag{1}$$

Classical RNN encoder–decoder implementations optimize this likelihood end-to-end, where the encoder computes a vector representation of the input, and the decoder generates tokens from left to right (Cho K. et al., 2014) [1].

Attention as a dynamic interface. To overcome the fixed-vector bottleneck, a context vector (or a set of context vectors) dependent on the generation step is introduced:

Table 1. **A brief chronology of key works that shaped the encoder–decoder paradigm**

Year	Work	Key Contribution to Encoder–Decoder
2006	Hinton, Salakhutdinov	Deep autoencoders as an "encoder–bottleneck–decoder" paradigm for representation learning (Hinton G. E., Salakhutdinov R. R. , 2006) [12]
2014	Cho et al.	Formalization of the RNN encoder–decoder as a conditional model $p(y \mid x)$ (Cho K. et al., 2014) [1]
2014	Sutskever *et al.*	Practical seq2seq with LSTM and "minimal assumptions" about seq→seq structure (Sutskever I. et al., 2014) [6]
2014 – 2015	Bahdanau *et al.*	Attention as a solution to the fixed-vector bottleneck; soft alignment (Bahdanau D. et al., 2014) [7]
2015	Luong *et al.*	Systematization of attention mechanisms (global/local) in neural machine translation (Luong T. et al.? 2015)[13]
2017	Vaswani *et al.*	Transformer encoder–decoder: self-attention + cross-attention without recurrence (Vaswani A. et al., 2017) [4]
2019 – 2020	Raffel *et al.*, Lewis *et al.*, Zhang *et al.*	Raffel et al., Lewis et al., Zhang et al. — Pretrained seq2seq as a universal task interface (T5/BART/PEGASUS) (Raffel C. et al., 2019) [8]
2020	Beltagy *et al.*	Longformer and LED as approaches to scaling seq2seq to long documents (Beltagy I. et al., 2020) [9]
2020 – 2021	Lewis *et al.*, Izacard & Grave	Retrieval-augmented generation (RAG) and fusion-in-decoder (FiD) (Lewis P. et al.. 2020) [10]
2022 – 2023	Alayrac *et al.*, Chen *et al.*, Li *et al.*	Modular VLMs with an explicit cross-modal "bridge" (Alayrac J.-B. et al., 2022) [14]

$$c_t = \sum_{i=1}^{T_x} \alpha_{t,i}\, h_i, \quad \alpha_{t,i} = \mathrm{softmax}\big(e_{t,i}\big), \tag{2}$$

where h_i are encoder states, and $e_{t,i}$ is a compatibility score between the decoder state at step tand position iin the input. This realizes a soft alignment between input and output (Bahdanau D. et al., 2014) [7].

Transformer: self-attention and cross-attention as a "contract". In the Transformer, the core attention operation is defined as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V, \tag{3}$$

and multi-head attention combines several such "heads." The encoder consists of self-attention layers (over the input), while the decoder includes masked self-attention (causal masking) and cross-attention to encoder outputs. Cross-attention serves

as the explicit "bridge" between representation and generation (Vaswani A. et al., 2017) [4].

Positional information. Since self-attention is permutation-invariant in the absence of additional signals, sequential structure is introduced via positional encodings. In the original Transformer, sinusoidal positional encodings are used:

$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right), \qquad (4)$$

$$\mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right).$$

The inclusion of positional information is also crucial for theoretical results concerning the universality of Transformer-like models in the seq2seq setting (Vaswani A. et al., 2017) [4].

Computational complexity as the "cost" of universality. For standard full self-attention, both computational cost and memory scale quadratically with sequence length. For an encoder–decoder Transformer, a rough estimate is:

– encoder self-attention: $O(T_x^2)$;

– decoder masked self-attention: $O(T_y^2)$;

– cross-attention: $O(T_x T_y)$.

This quadratic scaling motivates modifications of attention topology (Longformer/ LED), algorithmic accelerations of exact attention (FlashAttention), and alternative sequence modeling backbones for long contexts (e.g., state-space models such as Mamba) (Beltagy I. et al., 2020) Fig. 2. [9].

Variations of the Pattern: From Autoencoders to Retrieval and Multimodality. The autoencoder is a special case of the encoder–decoder paradigm, where the objective is to reconstruct the input, while the variational autoencoder (VAE) introduces a probabilistic interpretation (with an inference network as the encoder and a generative network as the decoder). In language models, denoising objectives such as those used in BART conceptually revive the autoencoder logic (with a "corrupted" input and its reconstruction) (Hinton G.E., Salakhutdinov R. R., 2006) [12].

Long-context approaches (Longformer/LED) modify the attention mechanism to process thousands of tokens while preserving the encoder–decoder topology for generation over long documents (Beltagy I. et al., 2020) [9].

Retrieval-augmented schemes (RAG, FiD) introduce external memory: the retriever acts as an "encoder of knowledge access," while the seq2seq generator aggregates the retrieved passages. In this setting, practical "universality" is achieved through a hybrid of retrieval and generation, where cross-attention serves as the mechanism for fusing multiple information sources (Lewis P. et al.. 2020) [10].

Multimodal systems demonstrate that the encoder–decoder paradigm provides a robust mechanism for transduction across different types of signals: image captioning (vision –> text), end-to-end ASR (audio –> text), as well as modern vision–language models (VLMs), which are built around a modular "bridge" between visual representations and a language decoder (implemented via controlled cross-attention or compact querying/bridging components) (Vinyals O. et al., 2014) [11].

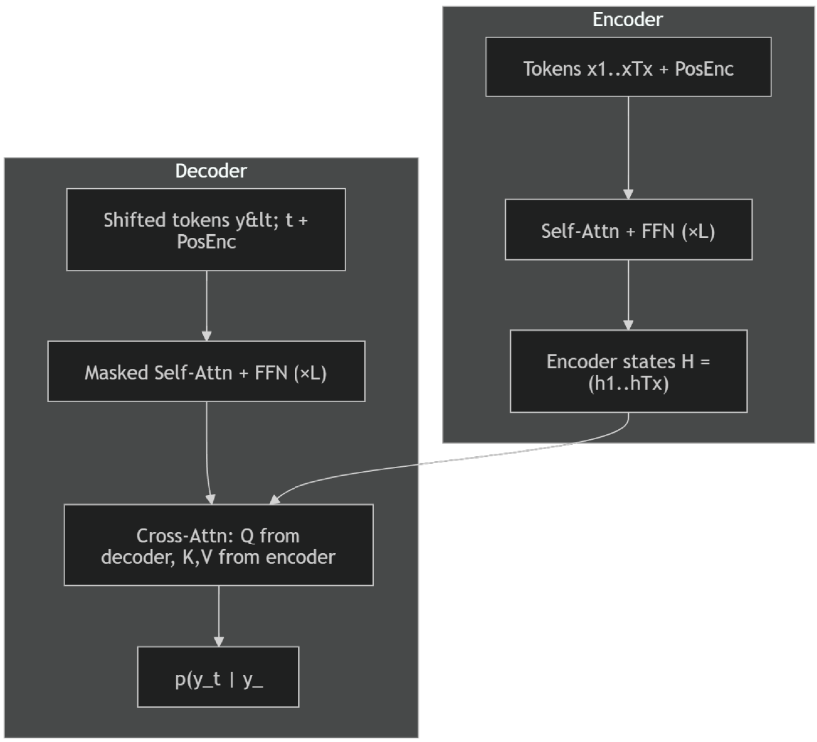

Fig. 2. **Generalized Transformer encoder–decoder architecture (network interface via cross-attention)**

Universality and Limitations. Universality as expressivity (theory). Formal statements about the "universality" of the Transformer family typically hold under specific conditions. In particular, it has been shown that Transformer models are universal approximators for certain classes of continuous sequence-to-sequence functions on compact domains; moreover, positional encodings remove permutation equivariance constraints and enable the approximation of arbitrary continuous seq2seq mappings on compact domains. Importantly, these results emphasize the distinct roles of self-attention and feed-forward layers in achieving such universality (Yun C. et al., 2019) [3].

However, theoretical expressivity does not guarantee practical universality: it does not address issues of sample efficiency, optimization stability, required data scale, pretraining regimes, or latent inductive biases. A historical example is the fixed-vector bottleneck: the limitation was not an inability of the network to represent the desired function in principle, but rather that such a factorization trained poorly and generalized inadequately for longer inputs—motivating the introduction of attention mechanisms (Bahdanau D. et al., 2014) [7].

Universality as an engineering contract (practice). Even as the decoder-only paradigm dominates (through scaling of causal language models and the in-context interface), practical systems often reintroduce explicit encoders (e.g., retrievers, vision encoders, audio front-ends) and fusion mechanisms (cross-attention or its

variants). This suggests that "decoder-only universality" in many modern pipelines is partly superficial: the encoder–decoder architectural skeleton reappears at the level of system modularity, even if the external interface presents as a single decoder (Lewis P. et al.. 2020) [10].

Main limitations: context length and the cost of attention. Quadratic self-attention makes universality computationally expensive. Different mitigation strategies introduce distinct trade-offs:

– Longformer/LED modifies the dependency topology (sparse attention) to approximately and structurally preserve useful global interactions over long documents (Beltagy I. et al., 2020) [9];

– FlashAttention retains exact attention while optimizing execution (an IO-aware algorithm), enabling practical scaling of context lengths and faster training (Dao T. et al., 2022) [20];

– alternative backbones such as Mamba abandon attention as the core mechanism, claiming improved linear scaling with sequence length and competitive performance across modalities (Gu A., Dao T., 2023) [15].

Scaling economics. Results on compute-optimal training emphasize that architectural comparisons cannot be disentangled from the question of how many tokens a model processes and under what computational budget. In practice, this strongly influences which topologies appear more "universal" under given infrastructure constraints (for example, decoder-only models are often easier to scale during training, whereas encoder–decoder models may offer advantages in conditional tasks and inference efficiency under certain regimes) (Hoffmann J. et al.,2022) [16].

Practical Implications and Recommendations. Architectural comparison is best viewed as a choice of inductive bias and dependency interface rather than as a matter of "trend." In conditional generation, the encoder–decoder paradigm imposes a clear structural discipline: the encoder constructs a representation of the input (often bidirectional), the decoder generates the output in a causal manner, and cross-attention regulates access to the input memory. In contrast, in decoder-only models, these roles are entangled within a single causal context. This can be advantageous (through interface unification), but also limiting (due to a less explicit dependency structure for complex X –>Y tasks) (Lewis M. et al., 2019) Table 2 [17].

Practical rule: if the task is formulated as "rich input Xand structured output Y," the encoder–decoder paradigm is often the most natural choice; if the task is "prompt-based text continuation," decoder-only models are typically simpler for product deployment; if the task requires "source attribution and updatable factual knowledge," retrieval-augmented seq2seq becomes a strong baseline option (Lewis P. et al.. 2020) Figure 3. [10].

Conclusions. 1. The encoder–decoder paradigm should be understood as a universal architectural contract between representation and generation: it is consistently instantiated across language, multimodal, and retrieval-oriented systems through an explicit interface (attention / cross-attention) (Vaswani A. et al., 2017) [4].

2. "Universality" in the sense of expressivity has rigorous theoretical formulations for Transformer-like models (universal approximation of classes of seq2seq functions under certain conditions), but this theory does not replace considerations of data, optimization, and inductive biases (Yun C. et al., 2019) [3].

Table 2. **Comparison of typical architectures for sequence modeling tasks**

Architecture	Structure	Typical Tasks	Advantages	Limitations
RNN seq2seq (without attention)	encoder → fixed code → decoder	early NMT	simple contract	bottleneck for long inputs (Sutskever I. et al., 2014) [6]
RNN + attention	encoder states + attention → decoder	translation, captioning	dynamic access to input	slow training due to recurrence (Bahdanau D. et al., 2014) [7]
Transformer encoder–decoder	self-attention + masked self-attention + cross-attention	translation, summarization, multimodal transduction	parallelism, modularity, strong inductive bias for $X \to Y$	quadratic self-attention (Vaswani A. et al., 2017) [4]
Pretrained seq2seq (T5/BART/PEGASUS)	encoder–decoder + self-supervised objectives	universal text-to-text tasks	high transferability	high pretraining cost (Raffel C. et al., 2019) [8]
Encoder-only (BERT-like)	bidirectional encoder + task head	classification, extraction, ranking	efficient "understa-nding"	not designed for generation without additional components (Devlin J. et al., 2018) [5]
Decoder-only (GPT-like)	causal self-attention	open-ended generation, in-context learning	unified prompt→ completion interface	more expensive access to long inputs; less explicit $X \to Y$ inductive bias (Radford A. et al.,2019) [18]
Long-context seq2seq (LED, etc.)	sparse attention y (enc/dec)	long-document summarization	extended context lengths	quality/engineering trade-offs (Beltagy I. et al., 2020) [9]
Retrieval-augmented (RAG/FiD)	retriever + seq2seq generator	knowledge-intensive QA	knowledge updatability, evidence aggregation	complexity of retrieval pipelines and evaluation (Lewis P. et al.. 2020) [10]

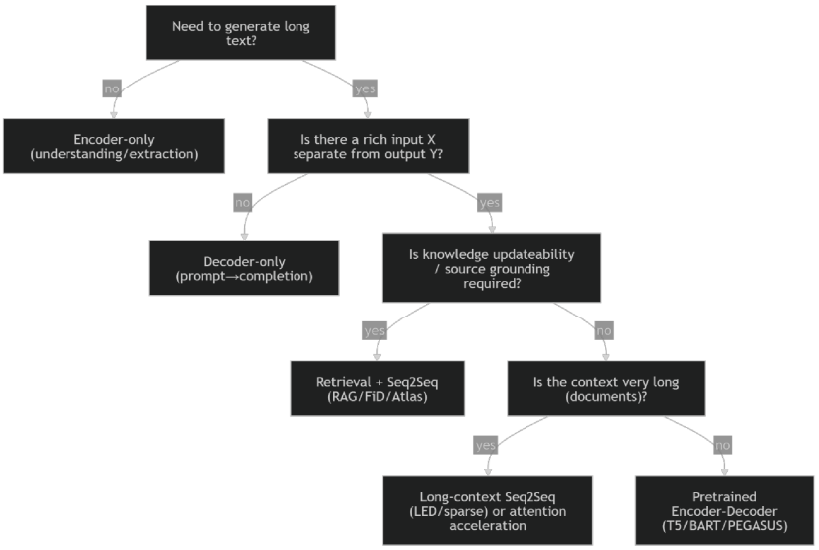

Figure 3. **Heuristic for architecture selection depending on task formulation (schematic)**

3. Practical universality is constrained by the cost of attention over long contexts; contemporary approaches evolve along three main directions: topological modifications of attention (Longformer/LED), algorithmic acceleration of exact attention (FlashAttention), and alternative sequence modeling backbones (e.g., Mamba) (Beltagy I. et al., 2020) [9].

4. The competition between encoder–decoder and decoder-only paradigms remain an open question: large-scale comparative studies indicate that encoder–decoder models can exhibit competitive scaling properties and inference advantages under certain regimes, even as decoder-only models dominate as the standard interface in many LLM scenarios (Zhang B. et al., 2025)[19].

1. Cho K. et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 2014. DOI: https://doi.org/10.48550/arXiv.1406.1078. URL: https://arxiv.org/abs/1406.1078

2. Brown T. B. et al. Language Models are Few-Shot Learners. 2020. DOI: https://doi.org/10.48550/arXiv.2005.14165. URL: https://arxiv.org/abs/2005.14165

3. Yun C. et al. Are Transformers universal approximators of sequence-to-sequence functions? 2019. DOI: https://doi.org/10.48550/arXiv.1912.10077. URL: https://arxiv.org/abs/1912.10077

4. Vaswani A. et al. Attention Is All You Need. 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762. URL: https://arxiv.org/abs/1706.03762

5. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. DOI: https://doi.org/10.48550/arXiv.1810.04805. URL: https://arxiv.org/abs/ 1810.04805

6. Sutskever I. et al. Sequence to Sequence Learning with Neural Networks. 2014. DOI: https://doi.org/10.48550/arXiv.1409.3215. URL: https://arxiv.org/abs/1409.3215

7. Bahdanau D. et al. Neural Machine Translation by Jointly Learning to Align and Translate. 2014. DOI: https://doi.org/10.48550/arXiv.1409.0473. URL: https://arxiv.org/abs/1409.0473

8. Raffel C. et al. Exploring the Limits of Transfer Learning with a Unified Text to Text Transformer. 2019. DOI: https://doi.org/10.48550/arXiv.1910.10683. URL: https://arxiv.org/abs/1910.10683

9. Beltagy I. et al. Longformer: The Long-Document Transformer. 2020. DOI: https://doi.org/10.48550/arXiv.2004.05150. URL: https://arxiv.org/abs/2004.05150

10. Lewis P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2020. DOI: https://doi.org/10.48550/arXiv.2005.11401. URL: https://arxiv.org/abs/2005.11401

11. Vinyals O. et al. Show and Tell: A Neural Image Caption Generator. 2014. DOI: https://doi.org/10.48550/arXiv.1411.4555. URL: https://arxiv.org/abs/1411.4555

12. Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks. Science. 2006. Vol. 313, № 5786. P. 504–507. URL: https://pubmed.ncbi.nlm.nih.gov/16873662/

13. Luong T. et al. Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2015. P. 1412–1421. DOI: https://doi.org/10.18653/v1/D15-1166. URL: https://aclanthology.org/D15-1166/

14. Alayrac J.-B. et al. Flamingo: a Visual Language Model for Few-Shot Learning. 2022. DOI: https://doi.org/10.48550/arXiv.2204.14198. URL: https://arxiv.org/abs/2204.14198

15. Gu A., Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. 2023. DOI: https://doi.org/10.48550/arXiv.2312.00752. URL: https://arxiv.org/abs/2312.00752

16. Hoffmann J. et al. Training Compute-Optimal Large Language Models. 2022. DOI: https://doi.org/10.48550/arXiv.2203.15556. URL: https://arxiv.org/abs/2203.15556

17. Lewis M. et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 2019. DOI: https://doi.org/10.48550/arXiv.1910.13461. URL: https://arxiv.org/abs/1910.13461

18. Radford A. et al. Language Models are Unsupervised Multitask Learners. OpenAI. 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

19. Zhang B. et al. Encoder-Decoder or Decoder-Only? Revisiting Encoder-Decoder Large Language Model. 2025. DOI: https://doi.org/10.48550/arXiv.2510.26622. URL: https://arxiv.org/abs/2510.26622

20. Dao T. et al. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. 2022. DOI: https://doi.org/10.48550/arXiv.2205.14135. URL: https://arxiv.org/abs/2205.14135

---

1. Cho K. et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 2014. DOI: https://doi.org/10.48550/arXiv.1406.1078. URL: https://arxiv.org/abs/1406.1078

2. Brown T. B. et al. Language Models are Few-Shot Learners. 2020. DOI: https://doi.org/10.48550/arXiv.2005.14165. URL: https://arxiv.org/abs/2005.14165

3. Yun C. et al. Are Transformers universal approximators of sequence-to-sequence functions? 2019. DOI: https://doi.org/10.48550/arXiv.1912.10077. URL: https://arxiv.org/abs/1912.10077

4. Vaswani A. et al. Attention Is All You Need. 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762. URL: https://arxiv.org/abs/1706.03762

5. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. DOI: https://doi.org/10.48550/arXiv.1810.04805. URL: https://arxiv.org/abs/1810.04805

6. Sutskever I. et al. Sequence to Sequence Learning with Neural Networks. 2014. DOI: https://doi.org/10.48550/arXiv.1409.3215. URL: https://arxiv.org/abs/1409.3215

7. Bahdanau D. et al. Neural Machine Translation by Jointly Learning to Align and Translate. 2014. DOI: https://doi.org/10.48550/arXiv.1409.0473. URL: https://arxiv.org/abs/1409.0473

8. Raffel C. et al. Exploring the Limits of Transfer Learning with a Unified Text to Text Transformer. 2019. DOI: https://doi.org/10.48550/arXiv.1910.10683. URL: https://arxiv.org/abs/1910.10683

9. Beltagy I. et al. Longformer: The Long-Document Transformer. 2020. DOI: https://doi.org/10.48550/arXiv.2004.05150. URL: https://arxiv.org/abs/2004.05150

10. Lewis P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2020. DOI: https://doi.org/10.48550/arXiv.2005.11401. URL: https://arxiv.org/abs/2005.11401

11. Vinyals O. et al. Show and Tell: A Neural Image Caption Generator. 2014. DOI: https://doi.org/10.48550/arXiv.1411.4555. URL: https://arxiv.org/abs/1411.4555

12. Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks. Science. 2006. Vol. 313, № 5786. P. 504–507. URL: https://pubmed.ncbi.nlm.nih.gov/16873662/

13. Luong T. et al. Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2015. P. 1412–1421. DOI: https://doi.org/10.18653/v1/D15-1166. URL: https://aclanthology.org/D15-1166/

14. Alayrac J.-B. et al. Flamingo: a Visual Language Model for Few-Shot Learning. 2022. DOI: https://doi.org/10.48550/arXiv.2204.14198. URL: https://arxiv.org/abs/2204.14198

15. Gu A., Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. 2023. DOI: https://doi.org/10.48550/arXiv.2312.00752. URL: https://arxiv.org/abs/2312.00752

16. Hoffmann J. et al. Training Compute-Optimal Large Language Models. 2022. DOI: https://doi.org/10.48550/arXiv.2203.15556. URL: https://arxiv.org/abs/2203.15556

17. Lewis M. et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 2019. DOI: https://doi.org/10.48550/arXiv.1910.13461. URL: https://arxiv.org/abs/1910.13461

18. Radford A. et al. Language Models are Unsupervised Multitask Learners. OpenAI. 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

19. Zhang B. et al. Encoder-Decoder or Decoder-Only? Revisiting Encoder-Decoder Large Language Model. 2025. DOI: https://doi.org/10.48550/arXiv.2510.26622. URL: https://arxiv.org/abs/2510.26622

20. Dao T. et al. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. 2022. DOI: https://doi.org/10.48550/arXiv.2205.14135. URL: https://arxiv.org/abs/2205.14135