RWKV Language Model

RWKV (pronounced RWaKuV) is an RNN with GPT-level large language model (LLM) performance that can be trained directly like a GPT Transformer (parallelizable).

RWKV combines the best features of RNN and Transformer: excellent performance, constant memory usage, constant inference generation speed, "infinite" context length, and free sentence embeddings. It is also 100% free of self-attention mechanisms.

The RWKV project was initially proposed by Bo Peng (Blink_DL), and as the project gained attention, it gradually developed into an open-source community.

On September 20, 2023, the RWKV open-source project officially joined the Linux Foundation. Today, the RWKV project is an open-source non-profit organization under the Linux Foundation, with some computing power previously supported by sponsors.

- Discord Forum
- HF Gradio-1 | RWKV-7-World-2.9B-v3
- HF Gradio-2 | RWKV-7-G1-0.4B

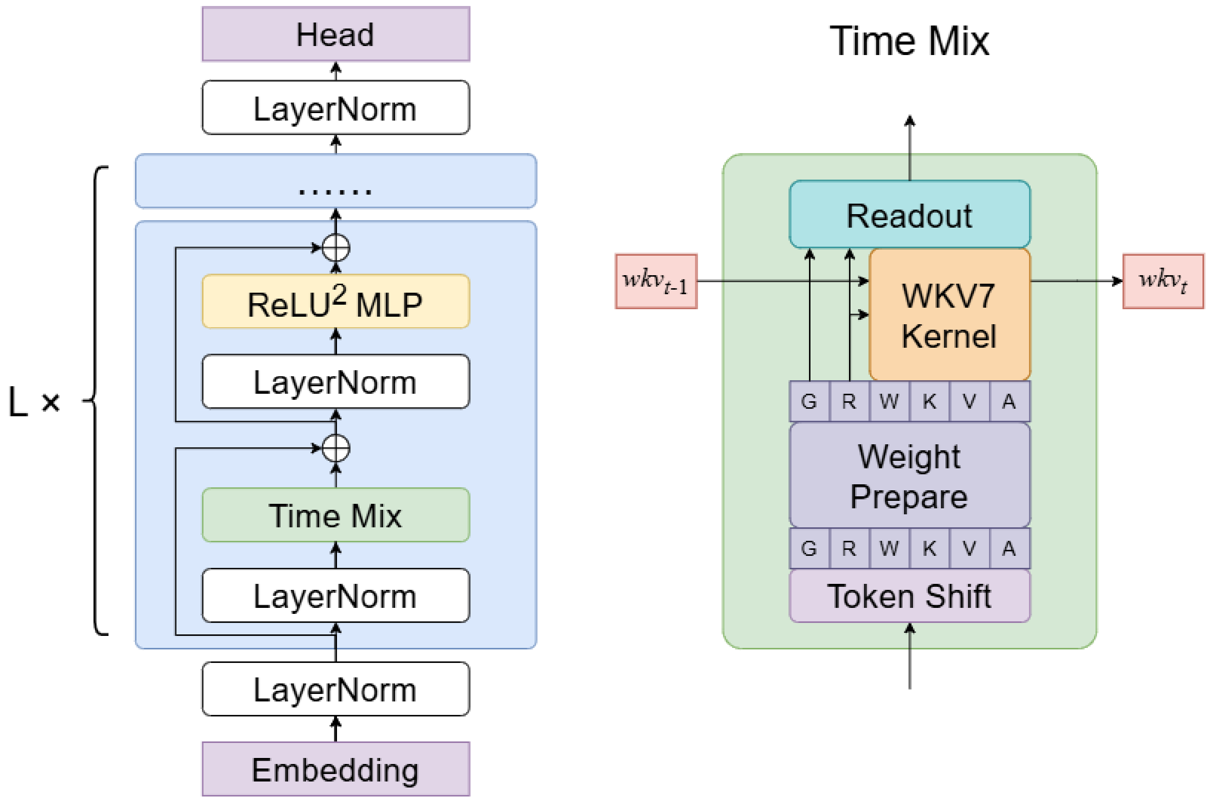
RWKV Architecture and Papers

RWKV-7 (Goose) is the latest version of the RWKV architecture. The paper was co-authored by Bo Peng and the RWKV community, published on March 18, 2025.

- **RWKV-7 Paper**: "RWKV-7 Goose with Expressive Dynamic State Evolution"
- **Paper Link**: arXiv:2503.14456

RWKV-7 adopts Dynamic State Evolution, surpassing the fundamental limitations of the TC0 expressive power of the attention/linear attention paradigm.

Click to view RWKV-7 Architecture Diagram



RWKV 5/6 (Eagle/Finch) architectures have several improvements based on the RWKV-4 architecture. Therefore, these two architectures are published in the same paper.

- **RWKV 5/6 Paper**: "Eagle and Finch: RWKV with Matrix-Valued States and Dynamic Recurrence"
- **Paper Link**: arXiv:2404.05892

RWKV-4 is the first official version of the RWKV model. The paper was co-authored by Bo Peng and the RWKV community and was first published on May 22, 2023. In October of the same year, the RWKV-4 architecture paper was accepted by EMNLP 2023.

- **RWKV-4 Paper**: "RWKV: Reinventing RNNs for the Transformer Era"
- **Paper Link**: arXiv:2305.13048

RWKV Model Version Status

RWKV has released open-source models of various parameter scales for each architecture version.

Version	**RWKV-V4**	**RWKV-v5-Eagle**	**RWKV-v6-Finch**	**RWKV-v7-Goose**	**RWKV-v7-G1**
Paper	Published	Published	Published	Published	Coming Soon

Version	**RWKV-V4**	**RWKV-v5-Eagle**	**RWKV-v6-Finch**	**RWKV-v7-Goose**	**RWKV-v7-G1**
Overall Status	**EOL**	**EOL**	**Stable**	**Stable**	**In Progress**
0.4B Model	Released	Released	No Plan	Released	Released
1.5B Model	Released	Released	Released	Released	📅 **Planned**
3B Model	Released	Released	Released	Released	📅 **Planned**
7B Model	Released	Released	Released	📅 **Planned**	📅 **Planned**
14B Model	Released	No Plan	Released	📅 **Planned**	📅 **Planned**

Which RWKV Models Should I Use?

Please use RWKV-7 series models. RWKV-7 models are based on the latest RWKV-7 architecture and latest datasets, therefore offering better performance.

Since RWKV-7 7B and larger models are still in training for 7B and larger parameter models, please use the RWKV-6-World-14B-V2.1 model; consider using the RWKV-6-World-7B-V3 model if your hardware cannot run the 14B model.

Tips

RWKV-7-World 7B/14B will replace the existing RWKV-6-World 7B/14B models once training is complete. Earlier RWKV versions have come to the end of their lifecycle, and existing models are only for archival purposes.

Differences Between RWKV and Transformer

- Advantages
    - Lower resource usage during runtime and training (VRAM, CPU, GPU, etc.).
    - **10 to 100 times lower computational requirements compared to Transformers with larger contexts**.
    - Supports linear scaling to any context length (Transformers scale quadratically).

  - Performs as well as Transformer architectures in terms of answer quality and generalization ability.
  - RWKV models' training data includes languages other than English (e.g., Chinese, Japanese, etc.), offering better multilingual capabilities than most existing open-source models.
- Disadvantages
  - RWKV base models are very sensitive to the format of prompts, and the format of prompts significantly affects the generation results.
  - Due to architectural design, RWKV models are weaker in **tasks requiring retrospection**, so prompts need to be appropriately ordered. For example, provide task instructions to the model first, then provide the material text needed to perform the task.

Basic Terminology of the RWKV Community

Concept	**Description**
RWKV	The model architecture itself, training code can be found here.
ChatRWKV	The official chatbot of RWKV (similar to ChatGPT but based on RWKV), code can be found here.
RWKV-4/5/6/7	Different architecture versions of RWKV. Note that using the latest RWKV-7 series models is recommended.
RWKV World	The base RWKV model trained with global languages, covering a broader and more diverse dataset, including training data in over 100 languages and some instruction training.
Raven	The official fine-tuned version of the RWKV-4 base model, including instruction training. However, since the RWKV-4 series is no longer updated, it is not recommended for continued use.
RWKV ABC/MIDI	RWKV music models based on ABC/MIDI format
RWKV CHNtuned / one-state-chat / role_play / novel ...	Fine-tuned models provided by the RWKV community, optimized for specific tasks or data types. Please prioritize using RWKV-7 series fine-tuned models.

RWKV Model Naming Rules

RWKV models typically have two naming conventions:

- RWKV-6-World-3B-v2.1-20240208-ctx4096.pth
- RWKV-x070-World-1.5B-v3-20250127-ctx4096.pth

The meaning of each field in the model name:

Field	**Meaning**
RWKV	Model name
6 / 070	RWKV model architecture, recommended to use RWKV-7 models
World	Model type, World indicates RWKV models trained with global languages, thus supporting multilingual tasks
3B / 1.5B	Model parameter scale, "B" stands for "Billions"
v2.1 / v3	Model training dataset version, v2.1 ≈ 1.1T , v3 ≈ 2.5T
20240208 / 20250127	Model release date
ctx4096	Pre-trained context length
.pth	RWKV model file format, also supports .gguf and .safetensors etc.

Prompting Format Guidelines

RWKV is a variant of RNN, and it is more sensitive to prompt formats compared to Transformer-based models.

RWKV is more suitable for two prompt formats: QA and Instruction.

QA Format

User: (Your question, e.g., "Please recommend three world famous novels suitable for five - year - old children.")

Assistant:

Tips

The QA (Question - Answer) format is the default training format for RWKV.

Here, User: represents the question asked by the user, and Assistant: represents the answer from the model. Therefore, we need to leave a blank after the last Assistant: to let the model continue writing.

Instruction Format

Instruction: Please translate the following Swedish into Chinese.

Input: hur lång tid tog det att bygga twin towers

Response:

Instruction: is the instruction given by the user to the model, Input: is the input provided by the user to the model, and Response: is the answer from the model.

Leave a blank after Response: to let the model continue writing.

Warning

Should not swap the positions of Instruction: and Input:

Due to the architectural design, RWKV has a relatively weak "recall" ability. If the RWKV model first receives the material content (Input) and then the instruction (Instruction), it may miss important information in the content when executing the instruction.

However, if you first tell the model **what instruction to execute** and then give the model **the input material content**, the model will first understand the instruction and then process the material content based on the instruction. Just like this:

Instruction: Summarize the following material text in one sentence.

Input: On February 22, 2025, the RWKV project held a developer conference themed "RWKV - 7 and Future Trends" in Caohejing, Shanghai, China. Developers, industry experts, and technological innovators from all over the country gathered together —— from well - known university laboratories to cutting - edge startup teams. The innovative energy on - site confirmed the excellent performance and far - reaching significance of RWKV - 7.

During the RWKV developer conference, 10 guests from academia, enterprises, and the RWKV open - source community brought in - depth sharing for developers, and the on - site audience interacted enthusiastically with the guests. For example, Yang Kaicheng from DeepGlint presented "RWKV - CLIP: A Robust Vision - Language Representation Learner", Hou Haowen from Guangming Laboratory presented "VisualRWKV: A Vision - Language Model Based on RWKV", Cheng Zhengxue from Shanghai Jiao Tong University presented "L3TC: Efficient Multimodal Data Compression Based on RWKV", Jiang Juntao from Zhejiang University presented "RWKV - Unet: Improving Medical Image Segmentation Results with Long - Distance Collaboration", etc.

During the conference, other AI enterprises also highly praised RWKV - 7, believing that it redefined the economic formula of AI infrastructure. The participants were also deeply touched by the demonstration of RWKV application results. Meanwhile, RWKV Yuanzhi Intelligence also shared RWKV - 7 and related demo presentations with thousands of developers at the 2025 Global Developer Conference.

Response:

Reference response:

The RWKV developer conference in Caohejing, Shanghai, China on February 22, 2025, attracted 10 guests from academia, enterprises, and the RWKV open-source community. The conference showcased the latest developments in RWKV-7 and its future trends. The on-site audience interacted with the guests and were impressed by the innovative energy on-site. The conference also featured presentations from other AI enterprises praising RWKV-7's redefinition of economic formula for AI infrastructure.