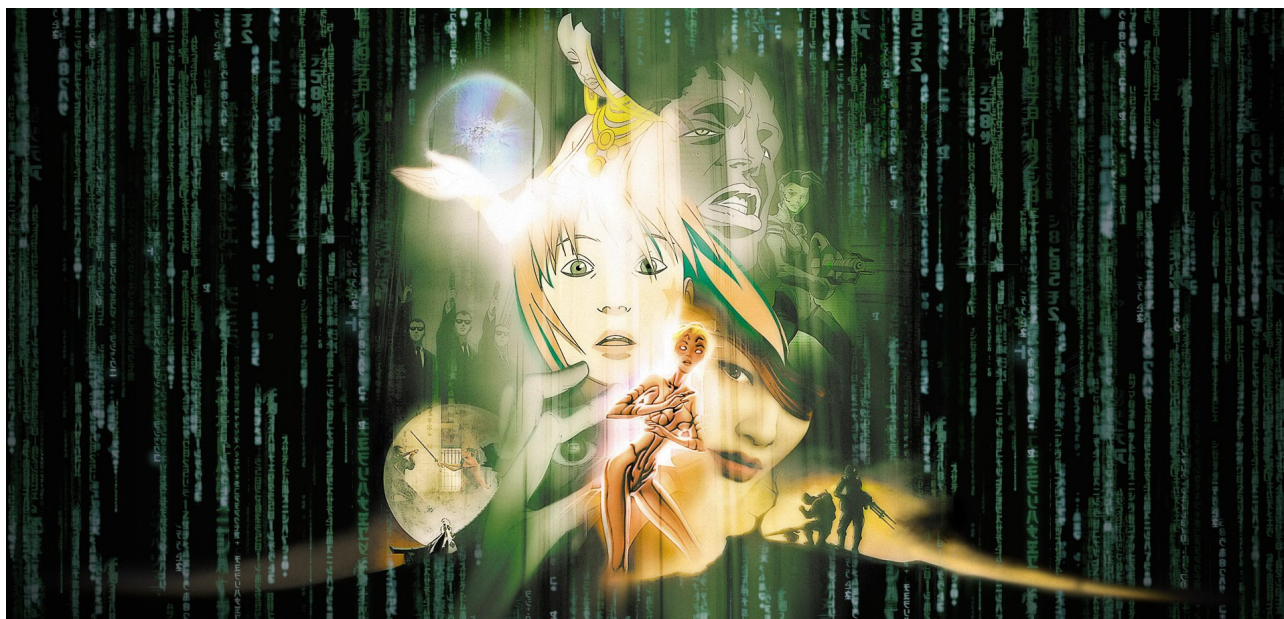


AI 分享

张驰

April 6, 2025



黑客帝国动画版（2003）海报

1 引子

近年，AI 对不少行业都形成了一定的影响，大家对 AI 的热情也愈发高涨。希望通过这篇文章以简单直观的方式介绍大模型领域的常见概念、各种时髦的名词及新兴技术（像是 AI Agent、Function Call、MCP 等），帮助大家了解当前的热点技术与工具，当工作中涉及 AI 时，可以有一定的判断。这篇文章主要讲解大模型相关技术与工具的背景、功能与使用场景，对于这些背后的原理只做简单阐述，如有兴趣可以进一步阅读参考文献。

2 大模型的工作原理

在阐述大模型相关的技术与工具前，需要大家首先对大模型的工作原理有一定了解。对于大模型的详细工作原理可参阅文献 [1]。其实大模型的工作原理很简单，就是四个字“文字接龙”，大模型根据输入的 Prompt 计算出下一个可能的字或词，再用此 Prompt 附加上新产生的字或词作为大模型新的输入，重复此过程直至得到完整的结果，大模型的简易工作原理如图1所示：

举个例子，如果向大模型输入 Prompt “今天天气怎么样？”，大模型大概会按以下流程进行工作并输出结果：

1. 分词，大模型首先对 Prompt 进行分词处理，将其分解为 token 序列“今天”、“天气”、“怎么样”与“？”，对于分词的详细介绍可参阅文献 [2]，另外 Hugging Face 也提供了一个 Tokenizer 的 Rust 实现 [3]；

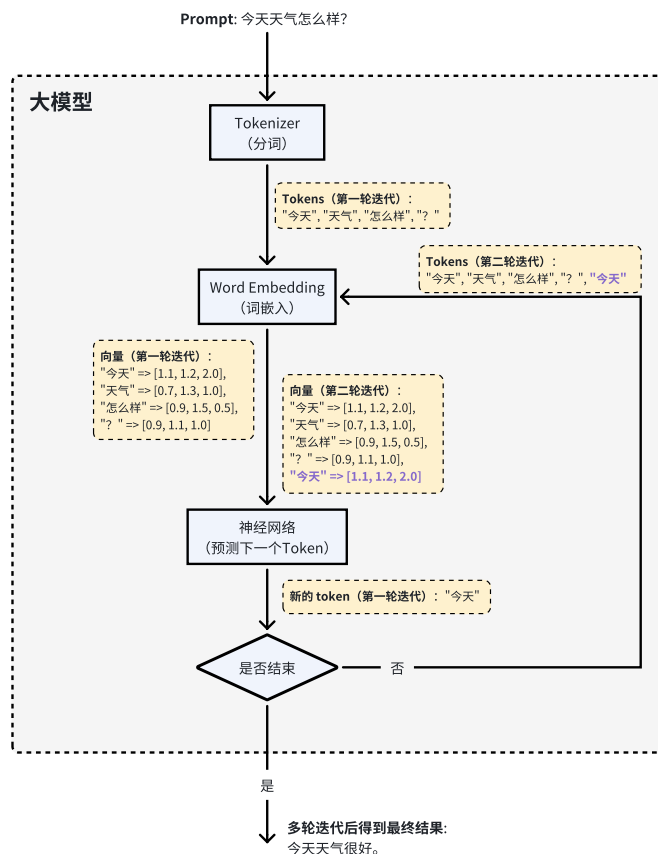


图 1: 大模型的工作流程

2. 词嵌入¹, 大模型随后对 token 进行词嵌入处理, 将每个 token 转化为一个向量, 比如“今天”被转化为向量 [1.1, 1.2, 2.0];
3. 预测, 大模型再将向量输入至神经网络进行计算得到一个新的 token, 如“今天”;
4. 附加, 大模型再将新的 token 附加至之前的 token 序列, 得到新的 token 序列“今天”、“天气”、“怎么样”、“?”与“今天”;
5. 重复步骤 2-4 直至得到最终结果。

引用

- [1] J. Alammam. “The illustrated transformer.” Accessed: 2025-04-03, Personal Blog. (2018), [Online]. Available: <https://jalammam.github.io/illustrated-transformer/>.
- [2] Glan 格蓝. “LLM 大语言模型之 Tokenization 分词方法 (WordPiece, Byte-Pair Encoding (BPE), Byte-level BPE(BBPE) 原理及其代码实现).” Accessed: 2025-04-04, Personal Blog. (2023), [Online]. Available: <https://zhuanlan.zhihu.com/p/652520262>.
- [3] [Online]. Available: <https://github.com/huggingface/tokenizers> (visited on 04/04/2025).

¹这里大家可能会有疑问, 这个功能是将 token 转化为向量, 那为什么不叫作“向量化”, 而是叫作“词嵌入”呢? “词嵌入”这个词最早出现于 Google AI 在 2013 年提出的 Word2vec 模型 [4], 词嵌入旨在将每个词赋予唯一的量化值, 并且语义较近的词的数量化值间的距离应当近, 语义较远的词的数量化值间的距离应当远, 另外每个词在不同的上下文中可能会有不同的语义, 为了承载语义的复杂性, 量化值通常是多维向量, 这里推荐阅读文献 [5]。

- [4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, 2013. arXiv: 1301.3781 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/1301.3781>.
- [5] 康斯坦丁. “没有思考过 Embedding, 不足以谈 AI.” Accessed: 2025-04-06, Personal Blog. (2023), [Online]. Available: <https://zhuanlan.zhihu.com/p/643560252>.