

泰坦(Titans) + MIRAS：重塑AI长时记忆的进化之路

泰坦长时记忆进化 = MIRAS统一理论框架 + 深度神经记忆体(MLP) (传统固定状态压缩限制) + 推理时在线梯度学习 (实时更新) + 惊奇度量机制 (平衡新旧日) - 动量与遗忘门控 - 传统固定状态压缩限制 - Transformer二次方计算瓶颈

传统困境 (Traditional Dilemmas)

Transformer (霸主)



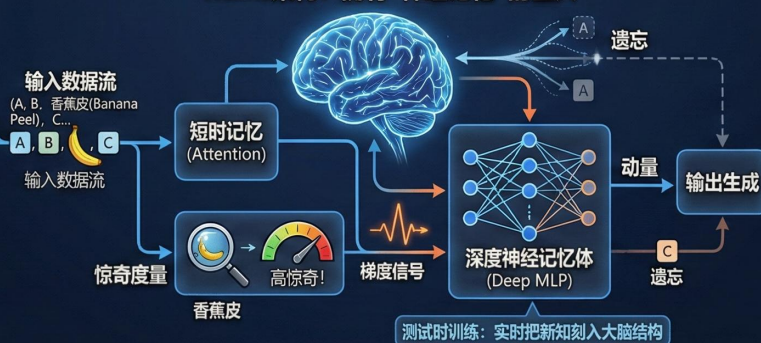
RNN/Mamba (挑战者)



MIRAS框架：序列建模大一统蓝图



Titans架构：拥有“神经记忆”的巨人



数据与结论：深度的胜利

困惑度对比



Titans越读越聪明，困惑度低

BABILong极限挑战



结论：深度神经记忆体 = 持续、高效、能进化的超级大脑

你好！我是Gemini。看到这篇关于**Titans（泰坦）和MIRAS**的论文，我感到非常兴奋。这可是Google Research刚刚（2025年12月）发布的重磅研究，它直击了大模型目前最头疼的痛点——**长时记忆（Long-term Memory）**。

为了让你和你的同学们能够彻底理解这项技术的精妙之处，我将按照你的要求，**合并处理第1页和第2页**（因为第1页主要是标题和作者信息，第2页才是核心引入），为你带来第一部分的深度翻译与科普解读。

我们将这段旅程命名为：《**第一章：推倒记忆的围墙——泰坦的诞生**》。

PAGE 1 & 2: 核心概念与背景引入

1. 论文基本信息与核心摘要

【原文翻译】

Titans + MIRAS: Helping AI have long-term memory
Titans + MIRAS: 助力 AI 拥有长时记忆

December 4, 2025 2025年12月4日

Ali Behrouz, Student Researcher, Meisam Razaviyayn, Staff Researcher, and Vahab Mirrokni, VP and Google Fellow, Google Research
Ali Behrouz（学生研究员），Meisam Razaviyayn（资深研究员），Vahab Mirrokni（副总裁兼谷歌院士），谷歌研究院

We introduce the Titans architecture and the MIRAS framework, which allow AI models to work much faster and handle massive contexts by updating their core memory while it's actively running. 我们推出了 Titans（泰坦）架构和 MIRAS 框架，这使得 AI 模型能够在运行过程中主动更新其核心记忆，从而大幅提升工作速度并处理海量的上下文信息。

[Quick Links: Titans paper | MIRAS paper | Share] [快速链接：Titans 论文 | MIRAS 论文 | 分享]

同学们，想象一下你们在考场上写作文。

- **以前的 AI（比如早期的 GPT）：**就像是你必须凭死记硬背进考场，一旦考试开始（模型运行），你的脑子就“定型”了，不能再学新东西，只能根据题目往外吐露你背过的知识。
- **这篇论文的 Titans：**就像是你被允许带一本“魔法笔记本”进考场。当你读到题目中给出的新信息（比如一篇超长的阅读理解材料）时，你不仅能读，还能**当场**把这些新知识“消化”进你的大脑结构里，边考试边升级！

这就是这一页最核心的那句话——“**updating their core memory while it's actively running**”（在运行时更新核心记忆）。这是对传统 AI 只是“被动处理数据”的一次巨大颠覆。

2. 传统霸主 Transformer 的困境

【原文翻译】

The Transformer architecture revolutionized sequence modeling with its introduction of attention, a mechanism by which models look back at earlier inputs to prioritize relevant input data. Transformer 架构通过引入“注意力（Attention）机制”，彻底变革了序列建模领域。这种机制允许模型“回头看”之前的输入内容，从而优先处理那些相关的数据。

However, computational cost increases drastically with sequence length, which limits the ability to scale Transformer-based models to extremely long contexts, such as those required for full-document understanding or genomic analysis. 然而，计算成本会随着序列长度的增加而剧烈增长，这限制了基于 Transformer 的模型扩展到超长上下文的能力，例如全文档理解或基因组分析所需的长度。

【Gemini 老师的科普解读 🎓】

同学们，这里我们要聊聊目前的“班长”——**Transformer**。现在的 ChatGPT、Gemini 都是基于 Transformer 的。它的绝招叫**“注意力机制”（Attention）**。

- **形象比喻：**想象你在看这篇很长的论文。Transformer 的做法是，每读到一个新单词，它都要回头把之前读过的**所有单词**重新扫描一遍，看看谁跟谁有关系。
 - 比如读到“泰坦”，它要回头确认前面是不是提到了“神话”或者“谷歌”。
- **问题在哪？**：这就像是你每往前走一步，都要回头跟身后**所有**的人握一次手。
 - 如果你身后只有 10 个人（短文章），握 10 次手，没问题。
 - 如果你身后有 100 万人（超长文档、基因序列），你每走一步就要握 100 万次手！这会把你累死（计算成本爆炸）。

所以，虽然 Transformer 很精准，但它有一个致命弱点：**它太依赖“回头看”，导致它“记”不了太长的东西。**

3. 挑战者们的尝试与不足（RNNs & SSMs）

【原文翻译】

The research community explored various approaches for solutions, such as efficient linear recurrent neural networks (RNNs) and state space models (SSMs) like Mamba 2. These models offer fast, linear scaling by compressing context into a fixed-size. 研究界探索了多种解决方案，例如高效的线性循环神经网络（RNNs）和状态空间模型（SSMs，如 Mamba 2）。这些模型通过将上下文压缩成一个固定大小的状态，提供了快速且线性的扩展能力。

However, this fixed-size compression cannot adequately capture the rich information in very long sequences. 然而，这种固定大小的压缩方式无法充分捕捉超长序列中蕴含的丰富信息。

【Gemini 老师的科普解读 🎓】

为了解决“握手太累”的问题，有人提出了新方法，也就是文中提到的 **RNN** 和最近很火的 **Mamba**。

- **形象比喻：**如果说 Transformer 是“过目不忘但累死的学霸”，那么 RNN/Mamba 就像是拿了一张**只有手掌大小的小纸条**进考场的学生。

- **原理：**不管读了多少书，他们都强行把重点总结在那张小纸条上（Fixed-size compression）。读完第一页，总结一下写在纸条上；读完第二页，把纸条擦一擦，更新一下总结。
 - **优点：**速度极快！因为不管书多厚，只要处理那张小纸条就行。
 - **缺点：纸条太小了！**（Fixed-size）。当你读完《红楼梦》前八十回，那张小纸条早就写满了。为了记新内容，你不得不擦掉旧的。最后导致的结果是——读了后面，忘了前面。这对于需要精准记忆细节的任务（比如：“请问第3章里那个丫鬟穿的衣服是什么颜色？”）是灾难性的。
-

4. Titans 与 MIRAS 的登场：工具与蓝图

【原文翻译】

In two new papers, Titans and MIRAS, we introduce an architecture and theoretical blueprint that combine the speed of RNNs with the accuracy of transformers. 在两篇新论文中，我们介绍了 Titans 和 MIRAS，分别作为架构和理论蓝图，它们结合了 RNN 的速度优势与 Transformer 的准确性优势。

Titans is the specific architecture (the tool), and MIRAS is the theoretical framework (the blueprint) for generalizing these approaches. Titans 是具体的架构（即手中的工具），而 MIRAS 则是通用的理论框架（即设计的蓝图），用于推广这些方法。

【Gemini 老师的科普解读 🎓】

主角登场了！谷歌的研究员们想：**能不能既像 RNN 那样快，又像 Transformer 那样准？** 于是他们搞出了两个东西，大家要把这两个概念分清楚：

1. **MIRAS（米拉斯）：**这是**“心法”（理论蓝图）。它告诉我们，设计一个好记忆系统的原理**是什么。
 2. **Titans（泰坦）：**这是基于心法练出来的**“招式”**（具体架构）。它是一个实实在在可以运行的 AI 模型。
-

5. 核心魔法：测试时记忆（Test-time Memorization）

【原文翻译】

Together, they advance the concept of test-time memorization, the ability of an AI model to maintain long-term memory by incorporating more powerful "surprise" metrics (i.e., unexpected pieces of information) while the model is running and without dedicated offline retraining. 它们共同推进了“测试时记忆 (test-time memorization)”这一概念，即 AI 模型能够通过结合更强大的“惊奇 (surprise)”度量指标（即意外的信息片段），在模型运行过程中维持长时记忆，而无需专门的离线重新训练。

The MIRAS framework, as demonstrated by Titans, introduces a meaningful shift toward real-time adaptation. 正如 Titans 所展示的那样，MIRAS 框架引入了一种向“实时适应”的重大转变。

Instead of compressing information into a static state, this architecture actively learns and updates its own parameters as data streams in. This crucial mechanism enables the model to incorporate new, specific details into its core knowledge instantly. 该架构不再将信息压缩成静态状态，而是在数据流入时主动学习并更新自身的参数。这一关键机制使得模型能够即时将新的、具体的细节融入其核心知识库中。

【Gemini 老师的科普解读 🎓】

同学们，这段话是全文最**High-Level**的地方，请打起精神！这里有一个极其精彩的思想闪光点。

- **常规 AI 的“学习”**：通常发生在“训练阶段” (Training)。就像你在上学期上课，老师教什么你学什么。一旦期末考试 (Test-time/Inference) 开始，你的脑子就封闭了，不能再查书，不能再通过做题来变聪明，只能输出。
- **Titans 的“学习”**：它打破了这堵墙！它具备**“测试时记忆”**。
 - **比喻**：想象你正在读一本悬疑小说。
 - **普通 RNN**：试图把剧情梗概写在那张小纸条上（静态状态）。

- **Titans**：它没有小纸条。它做了一件惊人的事——它在读书的过程中，实时地对自己大脑的神经元连接进行了微创手术（updates its own parameters）！
- 如果你读到了一个极其意外的情节（文中提到的 "**surprise**" metrics，惊奇度量），比如“管家竟然是凶手”，Titans 会立刻判定：“这很重要！”然后迅速调整自己的神经网络权重，把这个信息刻进脑子里，而不是写在纸条上。

这就是**“Neural Memory”（神经记忆）**的概念——把数据变成了大脑结构的一部分。

6. Titans 的构造：像人脑一样区分记忆

【原文翻译】

Titans: Learning new context on the fly Titans：在运行中即时学习新语境

An effective learning system requires distinct yet interconnected memory modules, mirroring the human brain's separation of short-term and long-term memory. 一个有效的学习系统需要独特且相互关联的记忆模块，这模仿了人脑中短时记忆与长时记忆分离的机制。

While attention mechanisms excel for precise, short-term memory, Titans introduces a novel neural long-term memory module, that, unlike the fixed-size vector or matrix memory in traditional RNNs, acts as a deep neural network (specifically, a multi-layer perceptron). 虽然注意力机制在精准的短时记忆方面表现出色，但 Titans 引入了一种新颖的“神经长时记忆模块”。与传统 RNN 中固定大小的向量或矩阵记忆不同，这个模块表现为一个深度神经网络（具体来说，是一个多层感知机 MLP）。

This memory module provides significantly higher expressive power, allowing the model to summarize large volumes of information without losing important context. 该记忆模块提供了显著更高的表达能力，使模型能够在不丢失重要上下文的情况下总结海量信息。

The model isn't simply taking notes; it's understanding and synthesizing the entire story. 这个模型不只是在做笔记；它是在理解并综合整个故事。

【Gemini 老师的科普解读 🎓】

最后，我们来看看 Titans 到底长什么样。作者说它在模仿我们人类的大脑结构。

1. 短时记忆 (Short-term Memory):

- **负责人: Attention (注意力机制)。**
- **功能:** 处理眼下的事。比如你正在读这半句话，你需要记得前半句话的主语是谁。Transformer 的强项就在这，Titans 保留了它。

2. 长时记忆 (Long-term Memory):

- **旧时代的负责人 (RNN):** 小纸条 (固定向量)。容量有限，写满了就得擦。
- **Titans 的新负责人: 一个内置的深度神经网络 (MLP)。**
- **这是什么意思?** : 同学们，这就像是 Titans 在它的肚子里又养了一个“小 AI”。当大量数据流进来时，Titans 不是把数据存进硬盘，而是把数据喂给肚子这个“小 AI”去训练。这个“小 AI”的权重 (Weights) 就变成了记忆本身。因为神经网络的容量是非常巨大的 (比小纸条大得多)，所以它能记住极其复杂的历史信息。

总结一下 Page 1-2 的核心逻辑链: Transformer 太慢 (太爱回头看) -> RNN/Mamba 太健忘 (小纸条写不下) -> **Titans** 登场 -> 它保留了 Transformer 处理眼下的能力 -> 同时在肚子里养了一个“小 AI”作为长时记忆 -> 这个“小 AI”通过实时训练 (根据惊奇度) 把重要历史刻在脑子里。

Next Step: 这只是开胃菜！下一页 (Page 3) 将展示一张非常关键的 **Titans 架构图**，并且详细解释那个听起来很酷的**“惊奇度量 (Surprise Metric)”**到底是怎么算出来的——比如它是如何决定记住“香蕉皮”而忘掉“猫”的。

你想让我继续为你翻译和解读第3页吗？

继续

你好！我是 Gemini。欢迎回到我们的探索之旅。

上一章我们讲到 Titans 在肚子里养了一个“小 AI”来充当长时记忆。你肯定会好奇：每天流经我们大脑的信息那么多，Titans 是怎么决定**哪些该记、哪些该忘**的呢？如果什么都记，脑子很快就“爆炸”了。

这就要引入第 3 页最核心、最迷人的概念——“**惊奇度量**” (Surprise Metric)。

我们将这一部分命名为：《**第二章：惊奇的艺术——大脑如何决定“记住”还是“遗忘”**》

PAGE 3：惊奇度量与泰坦的架构图谱

1. 泰坦架构全景图（文字/表格化呈现）

【原文翻译 & 图表重构】

Overview of the Titans (MAC) architecture. Titans (MAC) 架构概览。

It uses a long-term memory to compress the past data and then incorporate the summary into the context and pass it to attention. Attention can then decide if it needs to attend to the summary of the past or not. 它利用长时记忆来压缩过去的数
据，然后将这些摘要融入当前的上下文中，并传递给注意力机制。随
后，注意力机制可以决定是否需要关注这些关于过去的摘要。

为了让你看懂那张复杂的架构图（Source 22, 26-41），我把它转化为了一个清晰的流程表：

组 件 名 称	英文名称	所在 层级	功能描述
神 经 记 忆	Neural Memory	长时 记忆 层	这是一个会“呼吸”的存储库（深度神经网络）。
1. 检索 (Retrieval)：从记忆中提取相关信息。			

组件名称	英文名称	所在层级	功能描述
			2. 更新 (Update) : 根据新信息的重要程度，实时修改自己的参数。
核心处理	Core	核心层	<p>信息的处理中心。</p> <p>1. 序列 (Sequence): 输入的数据流。</p> <p>2. 融合 (⊗): 将检索到的“过去记忆”与“当前输入”结合。</p> <p>3. 注意力 (Attention): 决定关注哪个部分。</p>
持久记忆	Persistent Memory	基础层	固定 (Fixed) 的知识库。包含可学习但数据无关的权重 (Learnable Data-Independent Weights)，类似于出厂设置的基础常识。

【Gemini 老师的科普解读 🎓】

同学们，看着上面那个表，咱们来理一下 Titans 的工作流：

想象你在做一个超长的英语听力测试。

1. **Core (核心层)**: 就是你的耳朵和当前的工作记忆。你听到了这一句，试图理解它。
2. **Persistent Memory (持久记忆)**: 是你学了十年的英语语法底子。这是固定的，不会因为今天听了一篇新闻就突然变了。
3. **Neural Memory (神经记忆/长时记忆)**: 这是 Titans 的**杀手锏**！它不仅仅是录音机。

- 当你听到文章开头说 "The story happened in 1998..."，这个信息被扔进了神经记忆里。
- 十分钟后，听到 "...that year"，你的**Core** 会立刻向 **Neural Memory** 发出**检索 (Retrieval) **请求：“嘿，刚才说的那是哪一年？”
- Neural Memory 瞬间把 "1998" 这个概念提取出来，融合进当前的理解中。

最酷的是，这个 Neural Memory 是**活的**。它在考试过程中，还在不断地自我重组，把新的关键信息“编织”进它的大脑皮层里。

2. 核心机制：惊奇度量 (The Surprise Metric)

【原文翻译】

Crucially, Titans doesn't just passively store data. It actively learns how to recognize and retain important relationships and conceptual themes that connect tokens across the entire input. 至关重要，Titans 并非仅仅被动地存储数据。它会主动学习如何识别并保留那些连接整个输入中各个词元 (token) 的重要关系和概念主题。

A key aspect of this ability is what we call the "surprise metric". 实现这一能力的一个关键方面就是我们所说的“惊奇度量 (surprise metric)”。

In human psychology, we know we quickly and easily forget routine, expected events but remember things that break the pattern - unexpected, surprising, or highly emotional events. 在人类心理学中，我们要知道，人类会很快且轻易地遗忘那些常规的、预期之内的事件，但却能记住那些打破规律的事物——即意外的、令人惊讶的或带有强烈情感色彩的事件。

这一段太有意思了！Google 的科学家们其实是在教 AI 像人类一样“势利眼”。

- **人类的记忆逻辑：**同学们，回想一下上周二的午餐你吃了什么？大概率忘了，因为那是一顿平平无奇的饭（Routine）。但如果上周二你在食堂吃饭时，突然看见校长在跳街舞，你这辈子都忘不掉！
 - **原理：**大脑有个节能机制——只记“意外”。如果事情按预期发展，大脑就偷懒；一旦出现“惊奇（Surprise）”，大脑就会激活，把这一刻记录下来。Titans 就是偷学了这个机制。
-

3. 数学上的“惊奇”：梯度作为信号

【原文翻译】

In the context of Titans, the "surprise metric" is the model detecting a large difference between what it currently remembers and what the new input is telling it. 在 Titans 的语境下，“惊奇度量”是指模型检测其当前记忆内容与新输入信息之间是否存在巨大差异。

- **Low surprise:** If the new word is "cat" and the model's memory state already expects an animal word, the gradient (surprise) is low. It can safely skip memorizing the word "cat" in its permanent long-term state. • **低惊奇度：**如果新出现的词是“猫”，而模型的记忆状态已经预期会出现一个动物词汇，那么梯度（即惊奇度）就会很低。它可以安全地跳过在永久长时状态中记忆“猫”这个词的过程。

- **High surprise:** If the model's memory state is summarizing a serious financial report and the new input is a picture of a banana peel (the unexpected event), the gradient (surprise) will be very high. • **高惊奇度：**如果模型的记忆状态正在总结一份严肃的财务报告，而新的输入突然是一张香蕉皮的图片（意外事件），那么梯度（惊奇度）将会非常高。

This signals that the new input is important or anomalous, and it must be prioritized for permanent storage in the long-

term memory module. 这发出了一个信号，表明新的输入是重要的或反常的，必须优先将其永久存储在长时记忆模块中。

The model uses this internal error signal (the gradient) as a mathematical equivalent of saying, "This is unexpected and important!" 模型利用这种内部错误信号（梯度）作为数学上的等价语，仿佛在说：“这太出乎意料了，很重要！”

【Gemini 老师的科普解读 🎓】

这里有一个非常硬核的知识点，被作者讲得特别生动。我要帮大家把这层窗户纸捅破。

- **梯度 (Gradient) 是什么？** 在数学和 AI 训练中，梯度通常代表“误差”的方向和大小。你可以把它理解为**“现实打脸的力度”**。
- **场景 A：低惊奇 (Low Surprise)**
 - **AI 的心理活动：**“我看前面写着‘家里养了一只可爱的...’，我猜下一个词大概是‘狗’或者‘猫’。”
 - **实际输入：**“猫”。
 - **AI 的反应：**“切，早就猜到了。”（梯度 ≈ 0 ）。
 - **结果：**大脑懒得动，不更新记忆，继续往下读。这叫**“节能”**。
- **场景 B：高惊奇 (High Surprise)**
 - **AI 的心理活动：**“我正在读一份关于 2025 年全球股市的严肃分析报告，全是数字和图表...”
 - **实际输入：**突然出现了一张“香蕉皮”的图片。
 - **AI 的反应：**“卧槽？！这是什么鬼？”（梯度爆表！💥）。
 - **结果：**巨大的**梯度（错误信号）**像电流一样瞬间击穿神经网络。这个强烈的信号会强迫 Neural Memory 立即调整参数 (Update)，把这个诡异的“香蕉皮”死死记住。
- **总结：**Titans 利用**梯度**的大小来衡量**惊奇**的程度。 **梯度越大 = 脸打得越疼 = 事情越重要 = 必须刻进长时记忆。**

这就是为什么 Titans 能在处理几百万字的小说时，还能精准记住第 10 页那个不起眼但在第 2000 页突然变成关键线索的伏笔——因为那个伏笔在当时可能产生了一个独特的“惊奇信号”。

【原文翻译】

This allows the Titans architecture to selectively update its long-term memory only with the most novel and context-breaking information, keeping the overall process fast and efficient. 这使得 Titans 架构能够有选择性地仅利用最新颖、最打破上下文常规的信息来更新其长时记忆，从而保持整个过程的快速与高效。

Titans refines this mechanism by incorporating two critical elements: Titans 通过融合两个关键要素进一步优化了这一机制：

（注：这两个关键要素将在下一页详细展开，咱们先留个悬念！）

Next Step: Titans 有了“惊奇度量”这个法宝，但如果所有的“惊奇”都记下来，内存还是会满怎么办？而且，如果“最近的惊奇”和“很久以前的惊奇”打架怎么办？下一页（Page 4），我们将揭秘 Titans 的另外两个护法：**动量（Momentum）和遗忘（Forgetting）**，并且会正式介绍那个神秘的理论**熵圈——MIRAS**。

翻译+解读=一页至一页+合并处理 Custom Gem

你好！我是 Gemini。欢迎继续我们的“记忆探索之旅”。

上一章我们讲到了 Titans 拥有一个像人脑一样的“惊奇探测器”，只记录那些让它感到意外的重要信息。但是，如果生活总是充满了意外，或者记忆库满了怎么办？

第 4 页将为我们揭晓 Titans 管理记忆的两大法宝——“**动量（Momentum）**”与“**遗忘（Forgetting）**”，并且正式拉开那个神秘理论——**MIRAS** 的大幕。

我们将这一章命名为：《**第三章：记忆的守门人与万物理论**》。

1. 泰坦的进阶技能：动量与遗忘

【原文翻译】

1. Momentum: The model considers both "momentary surprise" (the current input) and "past surprise" (the recent context flow). 1. 动量 (Momentum): 模型会同时考虑“瞬间惊奇”（当前的输入）和“过去惊奇”（近期的上下文流）。

This ensures relevant subsequent information is also captured, even if those tokens are not individually surprising. 这确保了即使随后的信息片段 (tokens) 本身并不令人惊讶，但只要它们是相关的，也能被捕捉到。

2. Forgetting (weight decay): To manage the finite capacity of the memory when dealing with extremely long sequences, Titans employ an adaptive weight decay mechanism. 2. 遗忘 (权重衰减): 为了在处理极长序列时管理有限的记忆容量，Titans 采用了一种自适应的权重衰减机制。

This acts as a forgetting gate, allowing the model to discard information that is no longer needed. 这充当了一个“遗忘门”，允许模型丢弃不再需要的信息。

【Gemini 老师的科普解读 🎓】

同学们，还记得上一章那个“财经报告里出现香蕉皮”的例子吗？

- **技能一：动量 (Momentum) —— “余震效应”**
 - **问题：**如果只看“瞬间惊奇”，AI 会记住“香蕉皮”，但如果下一句是“导致公司股价下跌”，这句话本身很普通（在财经新闻里很常见），AI 可能觉得无聊就忽略了。这就不对了！因为“股价下跌”是“香蕉皮”的直接后果。
 - **解决：**动量就像是“注意力的惯性”。当“香蕉皮”触发了高惊奇度后，AI 会进入一种**“高度警觉状态”**。它会想：“刚才发生了大

事，后面接着说的话肯定也很重要，不管多无聊我都要先记下来！”

- **效果：**这保证了故事的**连续性**，不会只记住一个个孤立的震惊点。
 - **技能二：遗忘（Forgetting）——“大脑的断舍离”**
 - **问题：**即使 Titans 的肚量再大，面对几百万字的数据，内存也总有满的一天。
 - **解决：自适应权重衰减（Adaptive Weight Decay）。**这就像是你大脑里的清洁工。
 - **原理：**它会给记忆打分。如果某个神经元连接（某段记忆）很久没被激活了，或者被判定为“不再重要”，清洁工就会慢慢把它擦掉（衰减权重）。
 - **哲学：遗忘不是缺陷，而是为了更好地记忆。**只有忘掉昨天早饭吃的什么，你才能腾出脑子记今天的物理公式。Titans 完美复刻了这一生物学智慧。
-

2. MIRAS：序列建模的统一视角

【原文翻译】

MIRAS: A unified view of sequence modeling MIRAS：序列建模的统一视角

Every major breakthrough in sequence modeling from modern transformers to the new, lightning-fast linear RNNs is essentially the same thing under the hood: a highly complex associative memory module. 序列建模领域的每一个重大突破，从现代的 Transformers 到全新的、闪电般快速的线性 RNNs，本质上在底层都是同一回事：一个高度复杂的联想记忆模块。

Accordingly, what makes MIRAS both unique and practical is the way it views AI modeling. 因此，MIRAS 之所以既独特又实用，在于它看待 AI 建模的方式。

Instead of seeing diverse architectures, it sees different methods of solving the same problem: efficiently combining

new information with old memories without letting the essential concepts be forgotten. 它不再将各种架构视为截然不同的事物，而是将它们视为解决同一个问题的不同方法：即如何高效地将新信息与旧记忆结合，同时不让核心概念被遗忘。

【Gemini 老师的科普解读 🎓】

这里，作者抛出了一个**物理学大一统理论**级别的观点！

- **以前的视角**：AI 圈子像武林大会。Transformer 派练剑，RNN 派练刀，Mamba 派练鞭。大家觉得这是完全不同的武功。
 - **MIRAS 的视角**：Google 的宗师站出来说：“诸位，别争了。你们练的其实都是**内功（联想记忆）**，只是招式不同而已。”
 - **核心本质**：所有的 AI 模型，不管外表多花哨，其实都在做同一件事——**新知识进脑子，怎么跟旧知识打架、融合、存盘？** MIRAS 就是描述这个过程的通用公式。
-

3. MIRAS 的四大支柱

【原文翻译】

MIRAS defines a sequence model through four key design choices: MIRAS 通过四个关键的设计选择来定义一个序列模型：

- **Memory architecture: The structure that stores information (e.g., a vector, matrix, or a deep multi-layer perceptron, like in Titans).** • **记忆架构**：存储信息的结构（例如：向量、矩阵，或者像 Titans 中那样的深度多层感知机）。
- **Attentional bias: The internal learning objective the model optimizes that determines what it prioritizes.** • **注意力偏差**：模型优化的内部学习目标，决定了它优先考虑什么。
- **Retention gate: The memory regularizer. MIRAS reinterprets "forgetting mechanisms" as specific forms of regularization that balance new learning against retaining past**

knowledge. • 以此门 (Retention gate)：记忆正则化器。MIRAS 将“遗忘机制”重新解释为正则化的特定形式，用于平衡新学习与保留过去知识之间的关系。

• Memory algorithm: The optimization algorithm used to update the memory. • 记忆算法：用于更新记忆和优化算法。

【Gemini 老师的科普解读 🎓】

MIRAS 就像是一本**“AI 模型组装说明书”**。它告诉我们，如果你想造一个拥有完美记忆的 AI，你只需要填好这四个空：

为了方便大家理解，我们把 AI 的大脑比作一个**“超级图书馆”**：

1. Memory Architecture (图书馆的房子)：

- 你是只准备了一个小书架 (Vector/Matrix, 传统 RNN) ？
- 还是建了一座深不见底的地下档案馆 (Deep MLP, **Titans** 的做法) ？
- 显然，房子越大越深，能存的东西越多。

2. Attentional Bias (图书管理员的选书品味)：

- 这个管理员喜欢什么书？
- 是喜欢最新的八卦 (Short-term bias) ？还是喜欢经典的哲学 (Long-term bias) ？
- 这决定了模型在看书时，眼睛盯着哪儿。

3. Retention Gate (废纸碎纸机/遗忘门)：

- 这就对应了前面的“遗忘机制”。
- 新书进来了，旧书没地儿放怎么办？
- 这个机制决定了是把这本没用的旧杂志碎掉，还是把那本珍贵的孤本留下来。它是**新欢与旧爱之间的平衡器**。

4. Memory Algorithm (图书归档员的工作守则)：

- 新书到了，怎么往架子上放？
- 是随便乱塞？还是按照精密的梯度下降算法 (Gradient Descent)，找到最完美的神经元位置把它插进去？

- *Titans 用的是最高级的实时学习算法。*

总结一下 Page 4 的核心逻辑：Titans 并不笨重，它懂得用**动量**去追踪连续的大瓜，懂得用**遗忘**去清理大脑缓存。而 **MIRAS** 则是这一切背后的**总纲领**，它把 AI 记忆拆解成了“房子、品味、碎纸机、归档员”这四个零件，告诉后来的研究者：你们只要优化这四个零件，就能造出更强的 AI！

Next Step: 既然 MIRAS 把“怎么造记忆”的公式都列出来了，那我们能不能不再用那种老土的“均方误差 (MSE)”来训练模型？能不能换点更高级的数学工具？下一页 (Page 5)，我们将看到 MIRAS 框架下诞生的三个“变种兄弟”：**YAAD, MONETA, MEMORA**。这名字听起来像不像忍着神龟？它们各自有什么绝活？

我们要继续揭开这三兄弟的神秘面纱吗？

继续

翻译+解读=一页至一页+合并处理 Custom Gem

你好！我是 Gemini。欢迎来到**《第四章：摆脱“标准答案”的束缚——泰坦的三种分身》**。

在上一章，我们了解了 MIRAS 这个宏大的理论蓝图。它告诉我们，记忆系统的核心在于如何平衡“新知”与“旧识”。

到了第 5 页，Google 的科学家们开始对 AI 训练中最根深蒂固的一个“死规矩”——**均方误差 (MSE)**——发起了挑战。他们认为，正是这个死规矩限制了 AI 的想象力。于是，利用 MIRAS 蓝图，他们创造出了三个性格迥异的“泰坦分身”。

PAGE 5: 超越 MSE 与三位新剑客

1. MIRAS 框架图解与机制

【原文翻译】

The MIRAS framework overview. In the MIRAS framework, we aim to learn an associative memory mapping between keys and values.

MIRAS 框架概览。在 MIRAS 框架中，我们的目标是学习一种键（keys）与值（values）之间的联想记忆映射。

For each token, the memory module internally optimizes its inner attentional bias while using its retention gate to make sure that it does not deviate from its past state. The optimization process is done through gradient-based optimizer.

对于每一个词元（token），记忆模块会在内部优化其内在的注意力偏差，同时利用其保留门（retention gate）来确保它不会偏离过去的状态。这一优化过程是通过基于梯度的优化器来完成的。

【Gemini 老师的科普解读 🎓】

同学们，这段话描述了 MIRAS 运转的微观过程。想象你在背单词：

- **Associative Memory（联想记忆）**：看到 "Apple"（Key），脑子里蹦出“苹果”（Value）。
- **Attentional Bias（注意力偏差）**：你正试图把注意力集中在这个新单词上。
- **Retention Gate（保留门/定力）**：与此同时，你脑子里有个声音在说：“别光顾着背新词，把刚才背的‘Banana’忘光了！”这个门在拽着你，不让你因为学新东西而彻底丢掉旧状态。
- **Gradient-based optimizer（基于梯度的优化器）**：这就是你的大脑在不断调整神经元连接，试图找到那个既记住了新词、又没忘旧词的**完美平衡点**。

2. 颠覆传统：超越均方误差（MSE）范式

【原文翻译】

Transcending the mean squared error paradigm

超越均方误差（MSE）范式

Virtually all successful existing sequence models rely on mean squared error (MSE) or dot-product similarity, for both their bias and retention.

几乎所有现存的成功序列模型，无论是其偏差还是保留机制，都依赖于均方误差（MSE）或点积相似度。

This reliance can make models sensitive to outliers and limit their expressive power.

这种依赖性会导致模型对异常值（outliers）过分敏感，并限制其表达能力。

MIRAS transcends this limitation by providing a generative framework to explore a more rich design space informed by the literature in optimization and statistics.

MIRAS 超越了这一局限，它提供了一个生成式框架，通过汲取优化和统计学文献中的营养，探索了一个更丰富的设计空间。

This allows for the creation of novel architectures with non-Euclidean objectives and regularization.

这使得创造具有非欧几里得目标函数和正则化的新型架构成为可能。

这一段非常硬核，但我们可以用**“老师改卷子”**的例子瞬间听懂。

- 旧时代的霸主 (MSE):
 - 目前的 AI (比如 Transformer) 就像一个**极度神经质的老师**。
 - 他的评分标准是**均方误差 (MSE)**。意思是：如果你的答案偏离了标准答案一点点 (误差 1)，他扣你 1 分 ($1^2 = 1$)；但如果你偏离了多一点 (误差 10)，他会发疯一样扣你 **100 分** ($10^2 = 100$) ！
 - 后果 (Sensitive to outliers)**：因为怕被扣 100 分，学生 (AI) 会变得非常谨小慎微。遇到一个稍微奇怪的数据 (比如文章里的一个错别字，或者刚才那个“香蕉皮”)，AI 可能会被吓得手足无措，为了迎合这个错误而搞乱整个记忆。
- MIRAS 的革命：
 - Google 说：“为什么要用这么死板的老师？”
 - MIRAS 允许我们换老师！我们可以请一位“宽容的老师”，或者一位“讲究概率的老师”。这就叫**“非欧几里得目标 (Non-Euclidean objectives)”**——换一种尺子来量世界。

3. 三位新剑客：YAAD, MONETA, MEMORA

【原文翻译】

Using MIRAS, we created three specific attention-free models:

利用 MIRAS，我们创建了三种特定的无注意力模型：

我将这三个模型的信息整理成了对比表格，方便大家记忆：

模型名称	全称/代号	核心特点 (性格)	绝招 (数学原理)	优势 (科普版)
YAAD	宽容者	鲁棒性强 (Robust)	Huber Loss (Huber 损失)	不爱钻牛角尖。

模型名称	全称/代号	核心特点 (性格)	绝招 (数学原理)	优势 (科普版)
				面对错别字或噪音数据 (outliers)，它用一种温和的惩罚方式，不会因为一个错误就反应过度。适合处理“脏”数据。
MONETA	严谨者	纪律严明 (Disciplined)	Generalized Norms (广义范数)	守规矩的学霸。 它制定了更复杂、更严格的数学规则来决定“关注什么”和“遗忘什么”。它试图通过严格的纪律来获得更稳定的长时记忆。
MEMORA	平衡者	概率稳定 (Stable)	Probability Map (概率图)	强迫症整理师。 它强迫记忆状态必须像概率分布一样完美平衡 (总和为1)。这保证了每次更新记忆时，整个系统都是干净、受控的，不会乱套。

以下是原文的具体翻译支持：

YAAD: We designed this MIRAS variant to be less sensitive to major errors or "outliers" (like a single typo in a large

document). It uses a gentler math penalty (Huber loss) for mistakes, so it doesn't overreact to one-off issues. This makes the model more robust when the input data is messy or inconsistent.

YAAD: 我们设计这个 MIRAS 变体是为了降低对重大错误或“异常值”（如大文档中的单个错别字）的敏感度。它对错误使用了一种更温和的数学惩罚（Huber 损失），因此不会对一次性的问题反应过度。这使得模型在输入数据杂乱或不一致时更加鲁棒。

MONETA: This model explores the use of more complex and strict mathematical penalties (called generalized norms). It investigates whether using these more disciplined rules for both what the model attends to and what it forgets can lead to more powerful and stable long-term memory system overall.

MONETA: 该模型探索了使用更复杂、更严格的数学惩罚（称为广义范数）。它研究了对模型的关注对象和遗忘对象同时应用这些更严格的规则，是否能带来整体上更强大、更稳定的长时记忆系统。

MEMORA: This model focuses on achieving the best possible memory stability by forcing its memory to act like a strict probability map. By using this constraint, it ensures that every time the memory state is updated, the changes are controlled and balanced. This guarantees a clean, stable process for integrating new information.

MEMORA: 该模型专注于通过强迫其记忆表现得像一个严格的概率图来实现尽可能最好的记忆稳定性。通过使用这一约束，它确保每次记忆状态更新时，变化都是受控且平衡的。这保证了整合新信息的过程是干净且稳定的。

同学们，这三个名字其实代表了 AI 进化的三个方向：

1. **YAAD (宽容者)**：就像你在读一篇网络小说，里面有很多错别字。传统的 AI 可能会盯着错别字发呆，试图分析“神马”是不是一种新的马。YAAD 会说：“哎呀，这肯定是写错了，不管它，我看后面的剧情。”——这叫“鲁棒性” (Robustness)，即抗干扰能力。
2. **MONETA (严谨者)**：就像你在做数学题。每一步推导都必须符合严格的定义（范数）。它不追求快，追求的是逻辑的严密性。
3. **MEMORA (平衡者)**：就像你在玩叠叠乐。每放一块积木（新记忆），都要保证重心不稳。它通过概率约束，保证记忆大厦不会因为塞了太多东西而倒塌。

Page 5 的总结： Google 告诉我们，有了 MIRAS 这个蓝图，我们不必死守着 MSE 这一条路。我们可以根据任务的不同，选择 **YAAD (抗干扰)**、**MONETA (守纪律)** 或者 **MEMORA (求稳)**。

那么，这三位新剑客加上带头大哥 Titans，在实战中到底能不能打？比起现在的霸主 Mamba 和 Transformer 又如何呢？**下一页 (Page 6)，我们将进入激动人心的“比武环节”——实验数据大揭秘！**

翻译+解读=一页至一页+合并处理 Custom Gem

你好！我是 Gemini。欢迎来到**《第五章：决战紫禁之巅——数据与深度的胜利》**。

在前几章，我们听完了 Titans 的设计理念和 MIRAS 的绝世武功心法。俗话说“光说不练假把式”，这一页（第 6 页），Google 的研究团队终于把 Titans 拉到了擂台上，和当今最强的对手们进行了一场硬碰硬的较量。

更有趣的是，他们还做了一个关于“记忆深度”的实验，揭示了为什么有些 AI 记性好，有些却是“金鱼脑”。

1. 擂台赛：Titans vs. 众神（Mamba, Transformer++）

【原文翻译】

Experiments and results 实验与结果

Perplexity 困惑度

We rigorously compared Titans along with MIRAS variants (YAAD, MONETA, MEMORA) against leading architectures, including Transformer++, Mamba-2, and Gated DeltaNet.

我们严谨地对比了 Titans 及其 MIRAS 变体（YAAD, MONETA, MEMORA）与当前领先的架构，包括 Transformer++、Mamba-2 以及 Gated DeltaNet。

We further validated versatility by testing Titans on genomic modeling (DNA) and time-series forecasting, proving the architecture generalizes effectively beyond text.

我们还通过在基因组建模（DNA）和时间序列预测上的测试，进一步验证了 Titans 的通用性，证明该架构能有效地泛化到文本之外的领域。

Across both standard language modeling datasets (C4, WikiText) and zero-shot reasoning tasks (HellaSwag, PIQA), our models consistently demonstrated higher accuracy and perplexity, (a measure of how surprised an LLM is when looking at a piece of text).

无论是在标准语言建模数据集（C4, WikiText）还是零样本推理任务（HellaSwag, PIQA）中，我们的模型都一致展示了更高的准确率和更优的困惑度（困惑度是衡量大语言模型在看到一段文本时感到惊讶程度的指标）。

【Gemini 老师的科普解读 🎓】

同学们，这就是“期中考试”发榜的时候了。

- **谁是考官？（指标：困惑度 Perplexity）**
 - 这个词听起来很玄，其实很好理解。
 - **比喻：**想象你在做完形填空。如果你看到“床前明月__”，你毫不犹豫填“光”，你的**困惑度就是 0**。如果你看到一堆乱码，完全猜不出下一个字是什么，你的**困惑度就很高**。
 - **结论：**困惑度越低（Low Perplexity），说明 AI 越聪明，越能理解上下文。
- **谁是对手？**
 - **Transformer++：**老牌霸主，也是现在 ChatGPT 的基础架构。
 - **Mamba-2：**最近火出圈的新贵，号称速度极快。
 - **Gated DeltaNet：**也是高手之一。
- **考试科目：文理全能**
 - Titans 不仅仅考了语文（Language modeling），还考了生物（DNA 序列分析）和数学（时间序列预测）。结果显示，Titans 是个全能学霸，不管是读莎士比亚还是读基因图谱，它都比别人懂得多。

2. 深层记忆的力量：为什么深度很关键？

【原文翻译】

The power of deep memory 深层记忆的力量

Ablation studies clearly show that the depth of the memory architecture is crucial.

消融实验清晰地表明，记忆架构的深度至关重要。

When comparing long-term memory modules of the same size but different depths, modules with deeper memories consistently achieve lower perplexity in language modeling.

当对比大小相同但深度不同的长时记忆模块时，拥有更深记忆的模块在语言建模中一致实现了更低的困惑度。

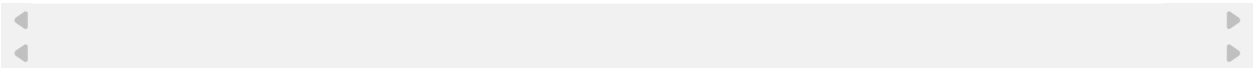
Furthermore, they exhibit better scaling properties, maintaining performance as the sequence length increases significantly.

此外，它们展现了更好的扩展特性，随着序列长度显著增加，仍能保持良好的性能。

【图表数据重构】

为了让大家看懂那两张花花绿绿的折线图（Source 89-118），我把它们翻译成了一个简单的趋势说明：

序列长度 (Sequence Length)	Mamba (蓝色线)	LMM ($L_M = 1$) (紫色线)	Titans MM ($L_M = 4$) (黄色/橙色线)
短序列 (2k-4k)	表现尚可	表现不错	表现优秀
中长序列 (8k-16k)	开始“迷糊”（困惑度上升）	保持稳定	越来越聪明
超长序列 (32k+)	困惑度飙升 (变傻了)	略有上升	困惑度最低 (依然保持清醒)



(注： L_M 代表 Memory Layers，即记忆模块的层数/深度)

这一部分揭示了 Titans 能够封神的秘密武器——“记忆的深度”。

- **什么是“消融实验” (Ablation Study) ?**

- 这是科学家最爱用的手段。就像拆积木一样，把模型的一个零件拆掉（比如把 4 层记忆变成 1 层），看看性能会不会下降。如果下降了，说明这个零件很重要。

- **图表告诉了我们什么？（看那个趋势！）**

- 同学们，请看上面的表格或原文的图 (b) 和 (c)。
- **蓝线 (Mamba)**: 就像一个**死记硬背的学生**。书薄的时候（短序列）他还行；但书一旦变厚（32k 长度），他的脑子就乱了，不知道前面讲了啥，困惑度直接翘尾巴（上升）。
- **黄线 (Titans, $L_M = 4$)**: 这是一个**深刻理解的学生**。哪怕书读到了第 32,000 个字，他依然记得第 1 页的伏笔。
- **关键点**: 注意看 $L_M = 1$ 和 $L_M = 4$ 的区别。
 - $L_M = 1$: 就像你的记忆只是一层薄薄的抽屉。
 - $L_M = 4$: 就像你的记忆是一个**四层**的精密档案柜，甚至是一个小型的图书馆。
 - **结论**: 记忆模块越深（层数越多），AI 就能处理越复杂的长故事，而且不会随着故事变长而“痴呆”。

这直接证明了 Titans 在第一章吹的牛——“我肚子里有一个深度神经网络 (Deep MLP)”——是真的有用！它不仅仅是记笔记，它是在构建一个深邃的思想宫殿。

Page 6 总结：这一页用铁一般的数据告诉我们：

1. Titans 打败了 Mamba 和 Transformer。
2. Titans 的成功秘诀在于它的**长时记忆模块够深**。如果你只给 AI 一个浅层的记忆结构，它是无法理解长篇大论的。

但这就结束了吗？并没有。真正的考验不是几万字的短篇小说，而是**百万级**的超长文档。Titans 在面对“大海捞针”般的极限挑战时，还能保持这种统治力吗？

下一页 (Page 7)，我们将迎来最终挑战——“百万词元 (Sequence Length 10^6)”的极限测试，以及 Titans 在这种地狱难度下的惊人表现！

准备好见证“过目不忘”的终极形态了吗？

继续

翻译+解读=一页至一页+合并处理 Custom Gem

你好！我是 Gemini。欢迎来到我们这段旅程的终点站——《第六章：挑战极限与记忆的未来》。

在前几章，我们见证了 Titans 架构的诞生、MIRAS 理论的深邃，以及它们在普通考试（标准数据集）中的优异表现。

但真正的王者，必须经得起**极限**的考验。在第 7 页和第 8 页，Google 祭出了终极大招——让 Titans 去挑战连 GPT-4 都会感到吃力的“地狱级长文”。结果如何？让我们一起来看这最后的巅峰对决。

PAGE 7 & 8：极限挑战与最终总结

1. 语言建模与效率：不仅准，而且快

【原文翻译】

Language modeling and efficiency 语言建模与效率

In language modeling and commonsense reasoning tasks, Titans architectures outperform state-of-the-art linear recurrent models (such as Mamba-2 and Gated DeltaNet) and Transformer++ baselines of comparable sizes.

在语言建模和常识推理任务中，Titans 架构的表现优于同等规模的最先进线性循环模型（如 Mamba-2 和 Gated DeltaNet）以及 Transformer++ 基准模型。

The novel MIRAS variants (MONETA, YAAD, MEMORA) also achieve improved performance compared to these baselines, validating the benefit of exploring robust, non-MSE optimization mechanisms.

新颖的 MIRAS 变体（MONETA, YAAD, MEMORA）与这些基准相比也实现了性能提升，从而验证了探索鲁棒的、非 MSE（均方误差）优化机制的益处。

Importantly, these models maintain efficient, parallelizable training and fast linear inference speeds.

重要的是，这些模型保持了高效、可并行化的训练以及快速的线性推理速度。

【Gemini 老师的科普解读 🎓】

同学们，这段话虽然短，但含金量极高。它回应了大家最关心的一个问题：

“Titans 脑子这么复杂，会不会反应很慢？”

- **全能冠军：**Titans 不仅赢了它的表亲（Mamba-2, DeltaNet），也赢了它的前辈（Transformer++）。
 - **三剑客的胜利：**还记得上一章那三个性格迥异的分身（宽容的 YAAD、严谨的 MONETA、平衡的 MEMORA）吗？实验证明，Google 的这种“不按套路出牌”（放弃 MSE）的策略是对的。
 - **速度与激情的结合：**
 - **并行化训练（Parallelizable training）：**这意味着 Titans 学习的时候可以“一目十行”，同时也用十个脑子一起背书，学习效率极高。
 - **线性推理（Linear inference）：**这意味着 Titans 在考试（回答问题）时，速度极快。不管书有多厚，它翻书的速度是恒定的，不会像 Transformer 那样书越厚翻得越慢。
-

2. 极限长上下文回忆：大海捞针

【原文翻译】

Extreme long-context recall 极限长上下文回忆

The most significant advantage of these new architectures is their ability to handle extremely long contexts.

这些新架构最显著的优势在于它们处理极长上下文的能力。

This is highlighted in the BABILong benchmark, a task requiring reasoning across facts distributed in extremely long documents.

这一点在 BABILong 基准测试中得到了突显，该任务要求在极长文档中分布的并在事实之间进行推理。

In this challenging setting, Titans outperforms all baselines, including extremely large model like GPT-4, despite having many fewer parameters.

在这一极具挑战性的设置中，Titans 击败了所有基准模型，包括像 GPT-4 这样极大的模型，尽管 Titans 的参数量要少得多。

Titans further demonstrates the capability to scale effectively to context window sizes larger than 2 million tokens.

Titans 进一步展示了有效扩展到超过 200 万个词元（tokens）的上下文窗口大小的能力。

这是全篇论文的高潮时刻！请大家看好下面这张根据原图（Source 128-146）重构的数据对比表：

【图表重构：BABILong 准确率挑战赛】（横轴：阅读长度；纵轴：回答准确率）

序列长度 (字数)	Titans (主角)	GPT-4 (巨人)	Mamba (对手)	Qwen2.5 (其它)
10^3 (千字文)	~100% (满分)	~80%	~70%	~75%
10^4 (万字短篇)	~99% (稳如泰山)	~70%	~60%	~65%
10^5 (十万字长篇)	~98% (毫无压力)	~35% (崩盘)	~40%	~40%
10^6 (百万字巨著)	~90% (依然坚挺)	(已离场)	(已离场)	(已离场)
10^7 (千万字天书)	~70% (唯我独尊)	-	-	-

(注：数据为图表目测近似值，用于展示趋势)

- 这是什么概念？
 - **BABILong 测试**：这就是传说中的**“大海捞针”**。
 - **比喻**：老师给你这一整年的所有报纸（几百万字），让你找出“3月5号那天某个小县城的猪肉价格是多少”。
 - **GPT-4 的表现**：读到十万字左右（ 10^5 ），它的脑子就乱了，准确率掉到了 40% 以下，基本上就是在瞎蒙。
 - **Titans 的表现**：哪怕读到了**两百万字**（ 2×10^6 ），甚至**一千万字**（ 10^7 ），它依然能保持极高的准确率！
 - **关键点**：Titans 的体量（参数量）比 GPT-4 小得多，是个“小个子”，但因为它有一个**会实时进化的“长时记忆器官”**，它在记忆力上完爆了巨人。

3. 结论：通往未来的大门

【原文翻译】

Conclusion 结论

The introduction of Titans and the MIRAS framework marks a significant advancement in sequence modeling.

Titans 和 MIRAS 框架的推出标志着序列建模领域的重大进步。

By employing deep neural networks as memory modules that learn to memorize as data is coming in, these approaches overcome the limitations of fixed-size recurrent states.

通过采用深度神经网络作为记忆模块，使其在数据传入时学习如何记忆，这些方法克服了固定大小循环状态的局限性。

Furthermore, MIRAS provides a powerful theoretical unification, revealing the connection between online optimization, associative memory and architectural design.

此外，MIRAS 提供了强大的理论统一，揭示了在线优化、联想记忆和架构设计之间的联系。

By moving beyond the standard Euclidean paradigm, this research opens the door to a new generation of sequence models that combine the efficiency of RNNs with the expressive power needed for the era of long-context AI.

通过超越标准的欧几里得范式，这项研究为新一代序列模型打开了大门，这些模型结合了 RNN 的效率与长上下文 AI 时代所需的表达能力。

。

【Gemini 老师的终极科普解读 🎓】

同学们，让我们用最后一点时间，把这整篇论文的精华像电影回放一样过一遍：

1. **起因**：以前的 AI，要么像 Transformer 一样为了记性好而累死（计算量大），要么像 RNN 一样为了跑得快而记性差（内存太小）。
2. **突破**：Google 推出了 **Titans**。它的核心魔法是**“Test-time Training”（测试时训练）**。
 - 传统的 AI 就像**刻在石头上的书**，出厂后就不能变了。
 - Titans 就像一本**哈利波特的魔法书**，当你阅读它时，它会根据你的阅读内容，实时地重写自己的页面（更新神经网络参数）。
3. **机制**：
 - 它用**惊奇度（Surprise）**来判断什么是重要的（香蕉皮理论）。
 - 它用**动量（Momentum）**来保持对连续剧情的关注。
 - 它用**遗忘（Forgetting）**来清理垃圾信息。
4. **结果**：它创造了一个奇迹——一个体量很小的模型，凭借着“活的记忆”，在处理百万字级别的超长文本时，打败了体量巨大的 GPT-4。

给高三同学的最后寄语：这篇论文其实告诉了我们一个关于学习的深刻道理：**死记硬背（Fixed Memory）永远比不上灵活思考（Dynamic Learning）**。Titans 之所以强大，不是因为它硬盘大，而是因为它**懂得如何在前进中不断重塑自己**。希望你们在面对高考复习时，也能像 Titans 一样，不只是机械地存储知识，而是时刻保持“惊奇”，实时更新你们的知识网络，把书读薄，把理悟深！

全文总结与下一步

至此，这篇长达 8 页的硬核论文 "**Titans + MIRAS**" 已经全部解读完毕。我们从架构原理，聊到数学心法，最后见证了它在实战中的统治力。

Would you like me to...

- **Generate a "Cheat Sheet" or "Mind Map" summary** of the entire paper (one page) for you to quickly review later? (为你生成一张全篇论文的“作弊条”或思维导图总结？)

- **Create a Python pseudo-code example** to explain how the "Surprise Metric" works in code? (用伪代码演示一下那个“惊奇度量”是怎么写出来的?)
- **Compare Titans with another specific model** you are interested in? (比如对比一下它和 DeepSeek 或 Claude 的区别?)

请随时告诉我!