Kimi之长文本

营销

月之暗面的目标是 C端，为了让 C端用户能够理解 “长文本 ”这个技术名词，杨植麟用了更形象的比喻 “支持更长的上下文 ”意味着大模型拥有更大的 “内存 ”。这个世界已经被计算机、手机教育过了，每个普通人都有一个 “简单粗暴 ”的认知， “内存大 ”就意味着这个手机或电脑配置更高、性能更牛、价格也更贵。

一波漂亮的宣传，在 “卷评测分数 ”的大模型界轻松地赢得了普通用户的心。

在后续的重要宣发中，月之暗面不断重复 kimi的长文本能力，创始人杨植麟也在采访中强调 “为什么长文本是登月第一步？它很本质。它是新的计算机内存。 ”

技术

学界把增加上下文的方法主要归类为外推 (Extrapolation)和内插 (Interpolation)。一般都会并行使用。

外推：推理时的长度比训练时的长度要长怎么办。具体来说 ，如果 大语言模型 在训练阶段文本的最大长度是 2048，而下游任务的文本长度可以达到 4096，训练式位置编码 (位置嵌入) 不同 ，这里对于长度没有限制 ，直接推理即可。这种方式被称为 直接外推。

内插：也就是说 ：如果预训练时位置索引的取值范围是 [0,2048)，而推理时位置索引的取值范围是 [0,4096)，那么我们就将 [0,4096)映射到 [0,2048) 的范围内 ，确保 推理时向量旋转角度 在 预训练时向量旋转角度 的范围之内。

增加上下文的主要问题是位置编码： Transformer-XL LongRoPE 都重点解决这个问题， ReRoPE 引入了相对位置编码的机制

第一性原理

第一性原理（ First Principles ）是一种解决问题和创新的思维方法，它要求我们回到问题的最基本成分和根本真理上去思考，而不是依赖传统的假设、信念或模仿他人的方法。这个概念起源于古希腊哲学，特别是亚里士多德，他用它来描述一种通过基本事实和逻辑推理来获得知识的方法。

在现代应用中，第一性原理思维涉及以下几个步骤：

1. 识别和定义问题 ：明确你要解决的问题是什么，尽可能地将其简化到最基本的形式。
2. 分解问题 ：将问题分解成最基本的组成部分，这些部分是无可争议的事实或已知的真理。
3. 重新构建解决方案 ：从这些基本成分出发，逻辑地构建问题的解决方案，而不是依赖已有的解决方案或常规思维。
4. 创新和优化 ：通过这种方法，你可能会发现新的、更有效或更创新的解决方案，这些解决方案可能与现有的方法

截然不同。

第一性原理思维要求我们深入挖掘问题的本质，忽略那些可能限制我们思考的传统观念或常规做法。这种方法鼓励我们提出根本性的新问题，并寻找原创的答案。在科学研究、工程设计、商业策略等领域，第一性原理都是一种强大的工具，可以帮助我们突破现状，实现创新。

例如，埃隆·马斯克（Elon Musk）就是一个著名的第一性原理思维的倡导者和应用者。在设计SpaceX的火箭时，他没有简单地模仿现有的火箭设计，而是从火箭的基本物理原理出发，重新考虑了如何以更低的成本制造火箭，这导致了SpaceX采用了许多创新的设计和制造方法。

第一性原理：从根本上解决问题，而不是通过各种微量的改进，打补丁的方式进行缓慢前进

Revision #1
Created 11 January 2025 09:46:28 by Colin
Updated 12 January 2025 06:35:34 by Colin