后训练扩展

简介

关于大语言模型在预训练阶段 scaling law (扩展定律) 的提法在 2020 年开始就已经提出并且受到关注[1], [2]。而在 2024 年底 OpenAI o 系列推理模型发布后又产生了 test-time scaling law 的说法，同时建立在此基础上的推理模型确实取得了可观的性能进步[3]。在此之外，关于智能体系统的扩展定律的说法也取得了相当的关注度[4]。很多时候它也可以被视作深度思考之外的另一种 test-time scaling 方法，在特定情况下可能取得比单纯的思维链扩展更好的性能表现[5]，同时近期流行的语言推理结合工具调用的交错式思维链某种经典的智能体工作方式和深度思考的结合[6], [7]。

但是与此同时，连接大语言模型大规模预训练和推理阶段的「后训练阶段」的扩展性反而较少被提及。事实上无论深度思考，亦或是复杂的智能体能力都显著依赖于后训练提供的能力。在大模型发展早期的 InstructGPT 当中就已经发现了少量的后训练就可以带来很强的人类偏好对齐和指令泛化能力[8]，因此在后续的两三年里大多数大模型的训练都遵循了大量预训练 + 少量后训练的方案，例如 DeepSeek-V3 当中真正后训练阶段的成本就只占到预训练的 0.2% 不到[9]。

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 Hours	2664K	199K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

Table 1: DeepSeek-V3 technical report 当中各训练阶段成本的估算 (假设单卡 H800 的租用价格是 $2 每小时)

但即使从来如此，这种做法也未必真正正确。因为从直觉上想，后训练理应取得足够的关注，这是由它们的训练目的决定的：

预训练 掌握语言的基本规律，学习文本的依赖关系。

中间训练 训练方式与预训练相同，但着重提升长上下文处理能力并使用高质量数据做技能退火。

后训练 学习运用语言的能力，包括规划、推理、指令遵循、执行动作等等。

DeepSeek-V3.2 的新方案

在备受关注的开源大模型 DeepSeek-V3.2 的 technical report[7] 当中提到了这样一句话：

Notably, this framework allocates a post-training computational budget exceeding 10% of the pre-training cost, unlocking advanced capabilities.

而另一方面，DeepSeek-V3.2 正式版在 continued pre-training 过程当中所使用的数据分布与 DeepSeek-V3.1 完全相同，仅用于启动 DeepSeek Sparse Attention 当中所使用的 lightning indexer，并且配合它适应稀疏注意力下的长上下文处理 (dense warmup stage & sparse training stage[7])。由此或许可以看到 DeepSeek-V3.2 的性能进步基本都来自于后训练的增强。

而在具体的实现上，DeepSeek-V3.2 训练当中所使用的后训练方案则相对复杂。大致分为以下几点：

- 专家蒸馏：在 DeepSeek-V3.2 base 的基础上针对数学、编程、通用逻辑推理、通用智能体任务、智能体编码、搜索智能体六个任务上分别训练出专用的语言模型，再使用它们合成特定领域的训练来蒸馏最终的大模型。
- 混合 RL：将推理、智能体和人类偏好对齐混合进同一个 RL 阶段当中，并应用大量改良后的 GRPO 算法进行训练，能够有效缓解灾难遗忘并且进一步实现能力提升。
- Speciale 模型的特殊处理：额外引入 DeepSeekMath-V2[10] 当中的训练数据和 reward model 进一步得到了推理增强的 DeepSeek-V3.2-Speciale 模型。

而为了实现这一切实际上工程上有很多的麻烦，为了同时照顾到各方面的通用能力，当前大模型的后训练阶段实现必然简单不了。

后训练扩展的可能未来

在此前也有一些广受关注的研究指出「大模型的知识与能力主要来自于预训练，而后训练过程仅仅是起到对齐或者唤醒知识的作用」[11]。但这里可能存在一个关键的因果倒置的逻辑错误，我们的思考方式或许不应该是「因为大模型能力主要来自于预训练，所以预训练效果决定了能力上限，需要更多关注预训练」，而是「因为此前预训练的算力和成本投入显著高于后训练，所以才会有大模型能力主要来自于预训练这样的结论」。事实上在很多新的研究里面后训练的贡献更多在帮助大模型学习行为模式或者语言应用技能上，而对于如今的大模型而言基础知识广度并不会成为瓶颈，此时学习新的行为模式就显得更为重要了。

但是当前大模型后训练阶段面对的任务复杂程度远超过单纯的语言建模，因此其实现方法上的复杂度也远超预训练阶段。因此针对后训练阶段的扩展可能就不仅限于在训练量和数据量上做简单增加，而可能更宽泛地指代对后训练阶段投入更多资源的行为。另一方面也不排除后训练阶段能够找到足以解决各种问题的统一方法的可能。

或许在未来我们会进入一个后训练规模超过预训练的时代，大语言模型就会更像一个独立的智能体，而非一个文本拟合的工具。很多现今存在的问题都可能由此得到解决，例如大模型的幻觉很大程度上就是由它的语言拟合本质所引发的，而足够强的后训练或许就可以有效缓解这些问题 (事实上从 GPT-2 时代至今这一问题已经改善许多了)。

预训练扩展的可能趋势

我们时常能听到「互联网上的人类数据已经快要用尽」这样的说法，这或许预示着大模型的预训练扩展已经快要走到尽头。尽管如此，很多新的预训练数据合成方法也被证明相当有效，例如 kimi k2 的 rephrasing[6]。另一方面纵观 DeepSeek-V3 系列的发展，我们也可能看到另一条可行的路线。

大模型预训练阶段通常需要使用几十 trillion 的 tokens，资源消耗通常非常巨大。另一方面从零开始进行预训练的话实际上会存在大量重复的消耗 (例如重新学习语言的基本规律以及各种世界知识)。DeepSeek 从 V3 开始直到 R1，V3.1，V3.2-Exp 以及 V3.2 都沿用了基本相同的模型结构主体，并且始终只使用了接续预训练 (mid-training 或 continued pre-training) 的方案，而没有再次预训练一个全新的 Base Model。但他们仍然用实践证明了，几乎相同的模型架构和预训练参数也可以做出能力的显著提升，如 DeepSeek-V3.2 就事实上具有接近 GPT-5 的 benchmark 表现[7]。

这可能带来的启发是，在未来或许从零开始预训练大模型的路线会逐渐减少，在同一个预训练大模型的基础上对结构做一些小调整、进行接续预训练做低成本更新的方案会成为主流。另一方面则是在预训练结果稳定之后，后训练扩展可能就会成为下一个重要的演进方向。

Bibliography

[1] J. Kaplan *et al.*, "Scaling Laws for Neural Language Models." Accessed: Dec. 13, 2025. [Online]. Available: http://arxiv.org/abs/2001.08361

[2] J. Hoffmann *et al.*, "Training Compute-Optimal Large Language Models." Accessed: Dec. 13, 2025. [Online]. Available: http://arxiv.org/abs/2203.15556

[3] OpenAI *et al.*, "OpenAI o1 System Card." Accessed: Mar. 23, 2025. [Online]. Available: http://arxiv.org/abs/2412.16720

[4] Y. Kim *et al.*, "Towards a Science of Scaling Agent Systems." Accessed: Dec. 13, 2025. [Online]. Available: http://arxiv.org/abs/2512.08296

[5] Y. Huang, Z. Tang, Z. Lin, P. Li, and Y. Liu, "Pessimistic Verification for Open Ended Math Questions." Accessed: Dec. 13, 2025. [Online]. Available: http://arxiv.org/abs/2511.21522

[6] K. Team *et al.*, "Kimi K2: Open Agentic Intelligence." Accessed: Aug. 04, 2025. [Online]. Available: http://arxiv.org/abs/2507.20534

[7] DeepSeek-AI *et al.*, "DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models." Accessed: Dec. 05, 2025. [Online]. Available: http://arxiv.org/abs/2512.02556

[8] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback.” Accessed: Mar. 24, 2025. [Online]. Available: http://arxiv.org/abs/2203.02155

[9] DeepSeek-AI *et al.*, “DeepSeek-V3 Technical Report.” Accessed: Feb. 24, 2025. [Online]. Available: http://arxiv.org/abs/2412.19437

[10] Z. Shao *et al.*, “DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning.” Accessed: Dec. 02, 2025. [Online]. Available: http://arxiv.org/abs/2511.22570

[11] C. Zhou *et al.*, “LIMA: Less Is More for Alignment.” Accessed: Dec. 14, 2025. [Online]. Available: http://arxiv.org/abs/2305.11206