

研究报告

(2024 年第 18 期 总第 50 期)

2024 年 12 月 18 日

LLM 时代小模型的应用潜力与挑战

清华大学五道口金融学院财富管理研究中心

张晓燕 张艺伟 张远远

【摘要】大语言模型在自然语言处理领域取得了显著成就。然而，由于其高昂的成本和资源消耗，其商业化应用仍处于早期阶段。相比之下，小语言模型凭借较低的训练成本和较小的计算资源需求等优势，展现出很大的潜力。本研究通过分析国内外小模型的发展现状与具体案例探讨了小模型的应用潜力与面临的挑战。研究发现：(1) 小模型体型虽小但功能强大，具有成本优势；(2) 我国小模型在专业领域中的表现能够超越国外通用模型，同时综合能力大幅提升，具有技术创新性；(3) 由于大模型对算

力资源的高需求，我国在资源与技术受限的情况下发展小模型，能够助力我国在全球 AI 竞赛中脱颖而出。本文为我国 AI 行业发展提供了战略方向。

Research Report

December 18, 2024

Application Potential and Challenges of Small Language Model in the LLM Era

Research Center for Wealth Management, PBC School of Finance,
Tsinghua University

Zhang Xiaoyan Zhang Yiwei Zhang Yuanyuan

Abstract: Large language models have made remarkable achievements in the field of natural language processing. However, due to their high cost and resource consumption, their commercial application is still in its early stages. In contrast, small language models have shown great potential due to their advantages such as low training cost and small computing resource requirements. This study explores the application potential and challenges of small models by analyzing the development status and specific cases of small models

at home and abroad. The study found that: (1) Small models are small in size but powerful in function, and have cost advantages; (2) my country's small models can outperform foreign general models in professional fields, and their comprehensive capabilities are greatly improved, which is technologically innovative; (3) Due to the high demand for computing resources by large models, my country's development of small models under the condition of limited resources and technology can help my country stand out in the global AI competition. This article provides a strategic direction for the development of my country's AI industry.

目录

一、 小模型的发展背景及意义.....	1
(一) 发展背景	1
(二) 研究意义	1
二、 LLM 时代小模型的发展现状	2
(一) 小模型定义	2
(二) 小模型与大模型比较	3
(三) 小模型的发展现状	4
1. 国际巨头深耕通用模型，中国企业聚焦多元垂直应用	4
2. 小模型目前的表现逐渐改善，然而距离大模型依然会有差距	6
3. 小模型体型虽小但功能强大	8
(四) 小模型的优势	9
1. 拥有巨大的成本优势	9
2. 指令微调以满足定制需求	11
3. 端侧搭配使用	12
4. 更低的延迟	13
三、 小模型的应用案例	13

(一) 国外小模型的应用案例.....	13
1. 通用领域小模型 Phi-3-Mini: 语言理解和生成.....	13
2. 通用领域小模型 OpenELM: 私有使用.....	17
(二) 国内小模型的应用案例.....	20
1. 垂直领域小模型智海·三乐: 智能教学.....	20
2. 垂直领域小模型夫子·明察: 法律咨询.....	23
3. 垂直领域小模型 DeepSeek-Coder-V2-Lite: 代码智能....	26
4. 垂直领域小模型 XuanYuan-6B: 金融咨询.....	29
(三) 中国小模型的优势.....	32
1. 在专业领域中的表现超越国外通用模型.....	32
2. 综合能力大幅提升.....	33
3. 具有技术创新性.....	37
四、 小模型面临的挑战及解决方案.....	39
(一) 处理复杂任务能力有限.....	39
(二) 数据质量依赖性.....	40
(三) 使用局限性.....	40
五、 小模型的未来发展趋势和建议.....	41

（一）发展小模型有助于提升我国在全球科技竞争中影响力.....	41
（二）中国发展小模型的建议.....	41
参考文献	43

一、小模型的发展背景及意义

(一) 发展背景

随着人工智能的迅速发展，以 ChatGPT 为代表的大语言模型的出现，在自然语言处理（NLP）领域取得了显著突破。这些大模型凭借其庞大的参数量和训练数据集，能够在多种任务中展现出令人瞩目的性能。然而，大模型虽然具备强大的处理能力，但其高昂的训练成本、复杂的部署要求以及巨大的资源消耗，限制了其在某些领域和场景中的普及和应用。

在此背景下，小语言模型应运而生。与大模型相比，小模型体积较小，计算需求低，训练成本显著降低，同时仍能够完成许多复杂的自然语言处理任务。这使得小模型在多个领域展现出了巨大的潜力和价值，成为人工智能领域的一股新兴力量。

(二) 研究意义

面对激烈的全球科技竞赛，我国在人工智能发展的路径上面临技术瓶颈和资源挑战。本文旨在通过案例分析和国内外比较，探索我国发展小模型的应用潜力与挑战，具有理论价值与实践意义：一是为小模型的研究与应用提供指导。通过分析国内外小模型的发展模式、性能表现及技术进步，为科研人员和企业提供了清晰的参考框架和有力的方向指引，推动其在实际应用中落地。

二是为政策制定与战略规划提供参考。在高端芯片受制约、全球竞争日益激烈的背景下，探索我国人工智能的发展路径，明确一条符合我国国情、兼顾效率与效益的发展之路，助力引导资源有效配置，优化产业布局。

二、 LLM 时代小模型的发展现状

(一) 小模型定义

大语言模型（Large Language Model, LLM）目前没有正式的定义，但通常是指由具有许多参数（数十亿个权重或更多）的人工神经网络组成，用于处理自然语言信息的大型人工智能模型。这些模型通过在大规模文本数据上进行训练，学习语言的语法、语义及上下文信息，从而能够理解和生成自然语言并执行多种任务。

小语言模型（Small Language Model, SLM）又被称作专业大模型（Specific Language Model, SLM），是指相对于大型语言模型而言，规模较小、参数数量较少，但在特定任务上表现出色的模型。尽管许多学者主要依据模型参数量与所利用训练数据规模来界定与评估何为大型语言模型，然而迄今尚未确立一个被广泛认可的临界标准。

(二) 小模型与大模型比较

大模型和小模型各自具有不同的特点和优势。表 1 从模型参数、训练数据、计算资源需求以及应用场景等角度，比较了大模型与小模型的异同。

表 1 小模型与大模型比较

	大模型	小模型
参数数量	通常在十亿的参数以上	参数数量较少，尚未确立临界标准
训练数据量	需要大量的训练数据来避免过度拟合	由于参数较少，可能需要的训练数据量也相对较少
计算资源需求	需要高性能硬件（如 GPU、TPU）、较长的训练时间、更高的计算成本	训练时间和计算成本相对较低，硬件需求也较低
应用场景	用于需要高精度和复杂决策的任务，如自然语言处理、图像识别、复杂的游戏 AI 等	适合实时应用、资源受限的环境或简单任务，如移动设备上的语音识别、智能家居控制、简单的预测任务等

资料来源：作者整理

从参数数量上来看，大模型通常指具有百万级、千万级甚至亿级参数的神经网络模型，如：GPT-3 为 175B、ERNIE 4.0 (百度) 为 100B 等。小模型相对于大模型参数数量较少，如：MobileLLM 为 125M、Phi-3 为 3.8B 等。

从训练数据量来看，大模型需要大量的训练数据来避免过拟合，ChatGPT-4 需要大约 13 万亿 tokens 的训练数据。小模型由于参数较少，需要的训练数据量也相对较少，Phi-3 的训练数据

仅为 3.3 万亿 tokens。

从计算资源需求来看，大模型需要大量的计算资源，如高性能 GPU 或 TPU。小模型计算资源需求较低，可以在普通的 CPU 上运行，适合在资源受限的环境中使用。此外，小模型的部署相对简单，适合在各种设备上运行，包括移动设备、嵌入式系统和物联网设备。

从应用场景来看，大模型通常用于需要高精度和复杂决策的任务，如自然语言处理、图像识别、复杂的游戏 AI 等。小模型更适合实时应用、资源受限的环境或简单任务，如移动设备上的语音识别、智能家居控制、简单的预测任务等。

(三) 小模型的发展现状

1. 国际巨头深耕通用模型，中国企业聚焦多元垂直应用

表 2 和表 3 例举了国内外近期发布的小模型，通过对比发现，像微软和 Google 这样的美国科技巨头专注于提升模型的通用性，致力于通用人工智能的发展路径。然而，这条赛道对于国内的中小企业来说，既“烧钱”又存在瓶颈，因为打磨出一款高水平的通用大模型需要投入大量的时间和资金。

我国的小模型在垂直领域的发展呈现出多元化的态势。除了以百度、阿里、华为、腾讯等巨头厂商深耕通用基础大模型研发，

更多的企业则根据自身产业生态的特点，开发专业化的垂直领域小模型。这类小模型不仅见效快，而且成本更低，为企业提供了灵活高效的发展路径。

表 2 国外发布的小模型

模型	发布机构	发布时间	模型参数量	类别
MobileLLM	Meta	2024 年 2 月	125M、350M、600M、1B 和 1.5B	通用
Phi-3	Microsoft	2024 年 4 月	3.8B、7B	通用
Gemma 2	Google	2024 年 6 月	9B	通用
SmoILM	HuggingFace	2024 年 7 月	135M、360M	通用
Llama 3	Groq	2024 年 7 月	8B	通用
GPT-4o mini	OpenAI	2024 年 7 月	未公布	通用
Mistral & Nvidia	NeMo	2024 年 7 月	12B	通用
DCLM	Apple	2024 年 7 月	7B	通用
xLAM	Salesforce	2024 年 7 月	1.35B 和 7B	通用

资料来源：作者整理

表 3 中国发布的小模型

模型	发布机构	发布时间	模型参数量	类别
Cornucopia(聚宝盆)	中国科学院成都 计算机应用研究所	2023 年 5 月	7B	金融
QiZhenGPT	浙江大学	2023 年 6 月	6B、7B、13B	医学
TransGPT	北京交通大学	2023 年 7 月	7B	交通
智海-录问	浙江大学、阿里	2023 年 8 月	7B	法律

	云、华院计算			
智海-三乐	浙江大学、高等教育出版社等	2023年8月	7B	教育
LightGPT	恒生电子	2023年10月	7B	金融
XuanYuan-6B	度小满	2024年3月	6B	金融
DeepSeek-Coder-V2- Lite-Base	DeepSeek	2024年6月	16B	代码
TechGPT 2.0	东北大学	2024年8月	7B	科研
MiniCPM3.0	面壁智能	2024年9月	4B	通用

资料来源：作者整理

中国的小模型更多地呈现出垂直化、应用化的特点，而国外小模型更多的是通用模型，这一现象背后的原因可能有：从成本与技术考量，国外科技巨头拥有强大的算力和数据资源，可以承担通用模型开发的高成本，同时通过多领域应用收回投入；国内以中小企业为主体，它们更倾向于使用低成本、高效率的垂直小模型，符合我国企业的“成本效益”策略。从政策与资源支持来看，国外更多资源集中于基础研究和技术创新，鼓励开发能够覆盖更多场景的通用模型；国内政策强调技术与实体经济的结合，支持行业定制化发展，为垂直小模型提供了良好环境。

2. 小模型目前的表现逐渐改善，然而距离大模型依然会有差距

得益于最新的 LLM pruning (剪枝)¹和 Knowledge distillation (知识蒸馏)²技术, SLM 目前的表现逐渐改善, 然而距离 LLM 依然会有差距。表 4 比较了 SLM 和 LLM 在通用、数学、推理、多语言处理任务中的表现, 每个模型的性能都是通过 Benchmark (标准的基准测试) 来衡量的, 分数越高代表性能越好。结果显示, SLM 取得了显著进步, 在一些任务中, SLM 得分已经接近一些 LLM。但是, SLM 在多个任务上的表现依然不及 LLM。

表 4 SLM 与 LLM 表现比较

		SLM				LLM			
Category	Benchmark	Llama 3.2 1B	Llama 3.2 3B	Gemma 2 2B IT (5-shot)	Phi-3.5 - Mini IT (5-shot)	Llama 3.2 11B	Llama 3.2 90B	Claude3 - Haiku	GPT-4o-mini
General	MMLU (5-shot)	49.3	63.4	57.8	69	73	86	75.2	82
Math	MATH (0-shot, CoT)	30.6	48	23.8	44.2	51.9	68	38.9	70.2
Reasoning	GPQA (0-shot)	27.2	32.8	27.5	31.9	32.8	46.7	33.3	40.2
Multilingual	MGSM (0-shot, CoT)	24.5	58.2	40.2	49.8	68.9	86.9	75.1	87

¹ LLM pruning (剪枝) 是一种通过去除模型中不必要的部分 (比如冗余的神经元或连接) 来减少模型体积和计算量的技术。

² Knowledge distillation (知识蒸馏) 是一种通过让一个小模型 (学生模型) 模仿一个大模型 (教师模型) 的输出, 从而提高小模型性能的技术。

资料来源：Meta Llama

3. 小模型体型虽小但功能强大

小模型体型虽小但表现优异。例如，微软发布的 Phi-3 模型是功能强大、具有成本效益的小语言模型（SLM），它在各种语言、推理、编码和数学基准测试中都优于相同大小和更高大小的模型。图 1 是微软 Phi-3 系列模型与其他小模型质量与规模的对比图。横轴表示模型参数的大小（以活动参数的亿级为单位），纵轴表示模型质量（根据“Massive Multitask Language Understanding”基准测试的分数）。可以发现，Phi-3 mini-4k 和 Phi-3 mini-128k 在较小参数情况下，性能已与 Llama3-8B 相当，甚至超过了一些 7-8B 的模型，如 Mistral 和 Gemma。除此之外，Phi-3 small 和 Phi-3 medium 在较小参数情况下能够达到更高性能。

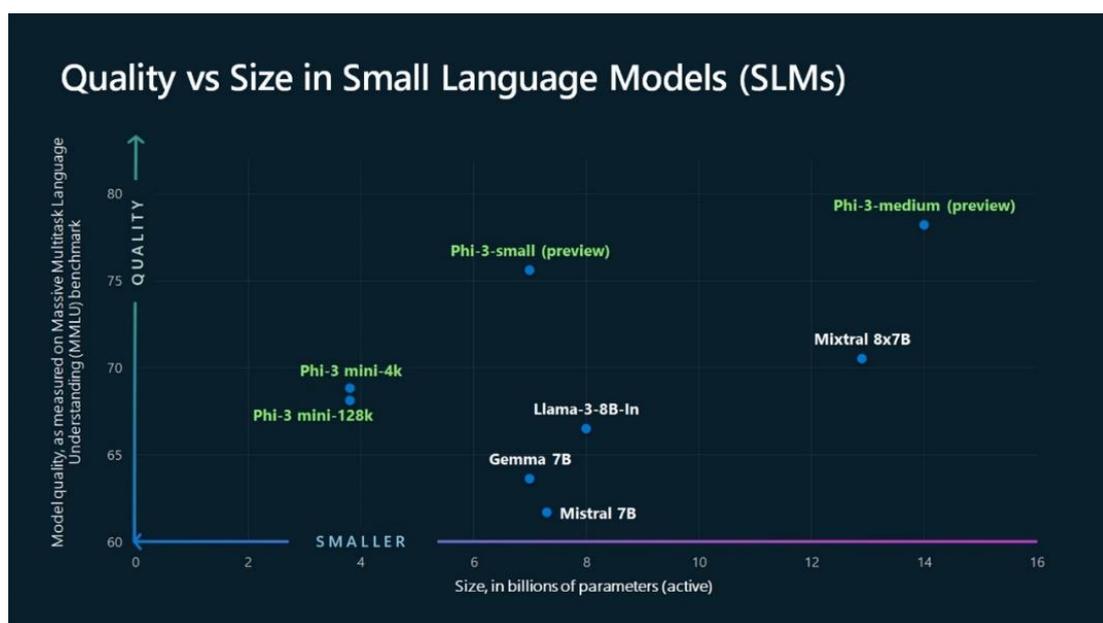


图 1 小语言模型的质量与规模

资料来源：Microsoft Phi

(四) 小模型的优势

1. 拥有巨大的成本优势

使用小模型的一个主要原因是其巨大的成本优势。以 GPT-4o mini 为例，其输入成本为 0.15 美元/百万 Tokens，输出成本为 0.6 美元/百万 Tokens，显著低于 GPT-4o 的输入成本 2.5 美元/百万 Tokens 和输出成本 10 美元/百万 Tokens。

尽管小模型的成本优势明显，但在模型能力上的表现与大模型相比差距并不大。图 2 是根据 OpenAI 的行业基准测试，比较了 GPT-4o mini 与一些大模型的表现。结果显示，GPT-4o mini 在 MMLU（大规模多任务语言理解）基准测试中的得分为 82%，虽然低于 GPT-4o 的 88.7%，但超过了 77.9% 的 Gemini Flash、73.8% 的 Claude 3 Haiku 和 69.8% 的 GPT-3.5 Turbo。同时，GPT-4o mini 在 MGSM（多语言小学数学数据基准测试）和 HumanEval（代码生成任务测试）中的得分分别为 87% 和 87.2%，也均高于 Gemini Flash 和 Claude Haiku，但略低于 GPT-4o。

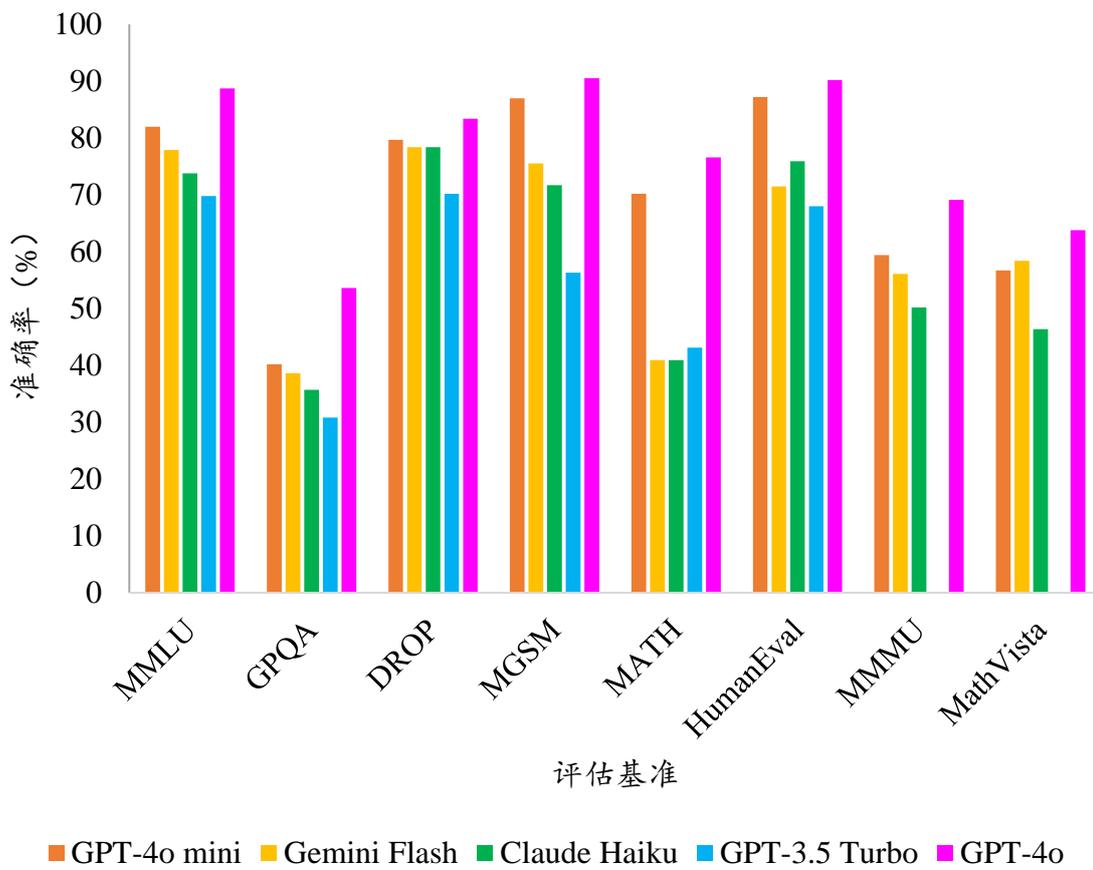


图 2 模型评估得分³

³ MMLU(Massive Multitask Language Understanding, 大规模多任务语言理解)

GPQA(Generalized Paragraph Question Answering, 广义段落问答)

DROP(Discrete Reasoning Over Paragraphs, 段落离散推理)

MGSM(Multi-Genre Sentence Matching, 多类型句子匹配)

MATH(Mathematics, 数学推理及解题能力)

HumanEval(代码生成)

MMMU((Massive Multitask Multimodal Understanding, 大规模多任务多模式理解)

MathVista(多模态数学推理)

资料来源: Open AI

小模型的精确率在某些基准测试中略低于大型模型,但这种差距并不显著。通过选择小模型,用户能够大幅降低计算成本,从而获得更高的性价比。

2. 指令微调以满足定制需求

大语言模型 (LLM) 适合需要编排涉及高级推理、数据分析和上下文理解的复杂任务的应用程序,意味着依赖其开发人员进行更新。而小语言模型 (SLM) 能够很好地执行简单的任务,用户可以在自己的服务器上运行它们,根据需求对其进行指令微调,拥有更多控制权。

例如,微软推出 **Phi-3** 系列小模型,使组织能够进行微调,自定义和扩展 AI 应用程序。**Phi-3-mini** 和 **Phi-3-medium** 模型现在支持无服务器微调,开发人员可以快速定制模型以适应云端和边缘场景,无需额外的计算资源。表 5 显示了 **Phi-3** 系列小模型在指令遵循和结构化输出方面做了显著的改进。结果表明, **Phi-3** 在更新后,指令遵循能力有所提升,其在各种应用场景中的表现也有所改善。

表 5 Phi-3 指令微调评测

	Phi-3-mini-4k	Phi-3-mini-128k
--	---------------	-----------------

基准（得分）	4.24 发布	6.24 更新	4.24 发布	6.24 更新
指令遵循	5.7	6.0	5.7	5.9
数据结构输出	11.5	52.3	11.9	60.1
段落问答	23.7	30.6	25.9	29.7
语言理解	68.8	70.9	68.1	69.7

资料来源：Microsoft

3. 端侧搭配使用

小模型关键创新之一是它能够在较小的设备（如手机、电脑、或其他移动设备、嵌入式系统等资源受限设备）上高效运行，而无需连接到云。这些设备的计算资源（如 AI 算力、内存等）相对有限，同时对端侧的能耗、发热等问题有着更为严格的要求。因此，小模型经过特别设计，以减小模型大小和优化模型架构，从而能够在端侧设备上高效运行。

以 **Apple Intelligence** 为例，这是苹果公司 2024 年推出的一款生成式 AI 小模型，它可以内置于 iPhone、iPad 和 Mac 等设备中，帮助用户完成写作、图像处理等任务。在写作辅助方面，**Apple Intelligence** 能够根据不同场景调整语气，校正语法和用词，生成重点摘要，并提供邮件回复建议。在图像处理方面，它可以根据文字描述生成新的表情符号、图片，或将草图转化为精美图像。此外，**Apple Intelligence** 还具备智能分类与摘要功能，并增强了 **Siri** 的能力，提供了更加自然的语言理解。

4. 更低的延迟

更少的参数意味着更少的处理时间。小模型更小，可以更快地处理请求，从而改善实时应用程序中的用户体验。例如，IBM Granite 3.0 版本推出的两款小模型：Granite 3.0 3B-A800M 和 Granite 3.0 1B-A400M，使用了“专家混合模型”（Mixture of Experts）。这意味着模型仅在推理时激活部分参数（分别为 800M 和 400M）来提高推理效率，减少计算资源的使用，适用于低延迟的应用场景。具体来看，Granite 3.0 1B-A400M 和 Granite 3.0 3B-A800M 的总参数计数分别为 10 亿和 30 亿，而它们在推理时的活动参数计数对于 1B 模型为 4 亿，对于 3B 模型为 8 亿。这使得这两个 SLM 都可以在提供高推理性能的同时最大限度地减少延迟。

三、小模型的应用案例

小型语言模型因其在训练成本低、计算效率高等方面的优势，其应用潜力正在逐步被重视。本文将介绍几款国内外具有代表性的小语言模型，探讨它们在特定任务中的表现，展示小模型在实际应用中的潜力。

(一) 国外小模型的应用案例

1. 通用领域小模型 Phi-3-Mini: 语言理解和生成

Phi-3-Mini 是微软推出的一款轻量级的语言模型，专为低资源、高效率、通用领域的任务而设计。Phi-3-Mini 拥有 3.8 亿参数和 3.3T tokens 的训练数据，在语言理解和生成中表现出色，能够与 GPT-3.5 媲美。

表 6 展示了 Phi-3-Mini 的功能和具体用途。Phi-3-Mini 在文本生成中，可以根据输入的提示词生成符合语境的文本内容；在文本补全中，可以根据部分输入内容预测和补全后续文本；在文本分类中，通过理解输入的语境和内容，将文本分类到预定义的类别中。在信息提取中，能够从非结构化文本中提取关键信息，如摘要生成、关键数据提取等等。在对话问答中能够根据输入问题，从给定的上下文或预训练知识中生成准确的回答。

表 6 Phi-3-Mini 功能和具体用途

功能	具体用途
文本生成	内容创作 ：生成博客文章、社交媒体文案、广告标题、产品描述等
	教育与培训 ：生成学习笔记、试题解析、课后总结等
	娱乐用途 ：生成短篇故事、对话剧本或创意写作内容
文本补全	代码补全 ：协助程序员快速完成代码段
	邮件与文档撰写 ：在用户输入关键内容后，自动补充完整的句子或段落
	报告总结 ：基于提供的概要内容，扩展出完整的文章
文本分类	情感分析 ：分析用户评论、社交媒体帖子是否为正面、负面或中性

	内容标记: 为文章或评论自动添加标签
	邮件分类: 将邮件分为“工作”、“社交”、“重要”等类别
文本翻译	语言翻译: 在常见语言之间翻译短文本
	语体转换: 将正式文档改写为口语风格, 或反之
信息提取	摘要生成: 从文章、报告中提取关键信息并生成摘要
对话问答	客户支持: 作为 FAQ 助手, 回答常见问题

资料来源: 作者整理

图 3 展示了 Phi-3-Mini 在实际应用中的卓越表现, 分为左、中、右三个部分。左侧部分, 用户提出任务: “写一首很难唱的短诗”。Phi-3-Mini 根据指令生成了一首复杂的诗作, 运用了绕口的短语和复杂的韵律结构, 使其在唱诵时具有很高的难度。中间部分, 用户询问: “在休斯顿 (Houston, TX) 有哪些适合 11 岁孩子的活动?” 模型推荐了两项活动: 参观休斯顿太空中心 (Space Center Houston) 和休斯顿动物园 (Houston Zoo), 并详细说明了这些地点的吸引点, 比如太空中心的互动展览和动物园丰富的物种展示。右侧部分, 用户提出要求: “为一篇论文起标题, 每个单词都以字母 P 开头, 论文主题是一个小型语言模型比大型模型更强大。” Phi-3-Mini 创造性地生成了标题: “Pinnacle Potential: Powering Progressive Paradigms with Pristine Predictive Prowess”。

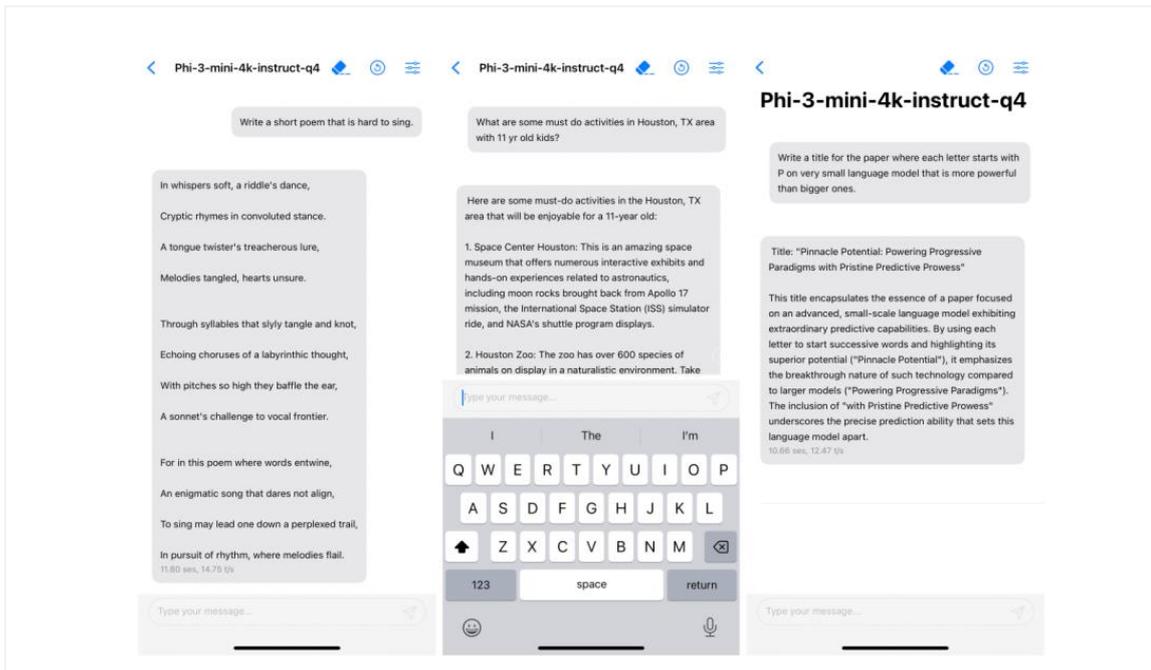


图 3 Phi-3-Mini 实际表现

资料来源：Phi-3-Mini

Phi-3-Mini 的创新之处在于使用独特的训练数据集——Phi-2 数据集的升级版，该数据集由精心筛选的公开网络数据及合成数据构成，规模庞大。训练数据精选自经过严格过滤的网络资源，以及由大型语言模型生成的合成数据。预训练过程分为两个独立且连续的阶段：第一阶段以网络资源为主，专注于培养模型的通用知识与语言理解；第二阶段则进一步筛选网络数据，并结合合成数据，强化模型的逻辑推理与特定领域的专业技能。此外，phi-3-mini 还经过优化，以确保其在鲁棒性、安全性及聊天格式适应性方面均达到卓越水平。

尽管 Phi-3-Mini 模型相对较小，但其在文本理解能力方面表

现非常强大，能够提供与大型模型相媲美的语言理解和推理能力。如表 7 所示，根据学术基准和内部测试的评估，Phi-3-Mini 在自然语言理解和推理中的表现超越 7-8B 参数的 Mixtral、Gemma 和 Llama-3-In，显示出其出色的文本理解能力。

表 7 Phi-3-mini 能力比较

Benchmark	Phi-3-mini 3.8B	Mistral 7B	Gemma 7B	Llama-3-In 8B
MMLU(大规模多任务语言理解)	68.8	61.7	63.6	66.5
HellaSwag(常识性自然语言推理)	76.7	58.5	49.8	71.1

资料来源：Microsoft

2. 通用领域小模型 OpenELM：私有使用

OpenELM 是苹果团队发布的一款通用小模型，包含了 2.7 亿、4.5 亿、11 亿和 30 亿四个参数版本。这款产品专为终端设备设计，聚焦隐私保护和数据安全，弥补了以往大规模语言模型 (LLM) 产品在这些领域的短板。

相较于微软的 Phi-3 Mini (3.8B 参数) 和谷歌的 Gemma (2B 参数)，OpenELM 以更低的运行成本脱颖而出，可直接部署于手机、笔记本电脑、平板等终端设备，无需依赖云服务器。

表 8 展示了 OpenELM 的使用场景和功能。OpenELM 能够

在移动设备上实现实时翻译、文本生成、个性化助手等功能，支持离线办公与学习。同时，其适用场景还包括可穿戴设备（如智能手表和 AR/VR 设备），可提供语音助手、健康提醒和会议字幕生成等服务。此外，它还能够在智能家居和车载系统等领域中广泛应用。

表 8 OpenELM 使用场景

设备	功能
移动设备(智能手机)	翻译工具: 提供实时离线翻译功能
	文本生成: 在移动办公场景中自动生成邮件、报告或会议纪要
	个性化助手: 集成到智能输入法中, 提供智能补全、修正和建议功能
可穿戴设备(可穿戴手表、AR/VR 设备)	语音助手: 支持用户通过语音查询天气、设置提醒或回复消息。
	健康助手: 根据用户语音输入生成健康建议或日程提醒
	AR/VR 设备: 提供自然语言交互, 例如在虚拟会议中生成实时字幕或摘要
智能家居设备	语音助手: 提升语音识别和回答的智能性, 无需连接云端, 增强隐私保护
	智能家电控制: 支持通过自然语言指令控制灯光、空调、电视等家电

汽车与车载设备	车载助手：提供离线语音导航，避免依赖网络；智能回复语音指令，例如控制音乐、调整空调
---------	---

资料来源：作者整理

这种面向终端运行的特性相较于大型模型具有显著优势：一是低延迟，在端侧设备运行，无需联网，响应速度快；二是隐私保护，数据无需上传云端，保护用户隐私；三是能耗优化，轻量模型适合嵌入式设备，运行高效。

OpenELM 使用了分层缩放策略，优化了 Transformer 模型各层之间的参数分配。这种方法不仅节省了计算资源，还提高了在相对较小数据量上的训练性能。表 9 显示了 OpenELM 与其他规模相近模型在平均准确率上的对比结果。与 OLMo 相比，OpenELM 的准确率提升了 2.36%，同时预训练所需的标记数量减少了一半。

表 9. OpenELM 对比 Public SLMs.

模型	公共数据集	公开		模型规模	预训练标记	平均准确率(%)
		代码	权重			
OPT	否	是	是	1.3B	0.2T	41.49
PyThia	是	是	是	1.4B	0.3T	41.83
MobiLlama	是	是	是	1.3B	1.3T	43.55
OLMo	是	是	是	1.2B	3.0T	43.57
OpenELM	是	是	是	1.1B	1.5T	45.93

资料来源：Apple

OpenELM 的推出，标志着小模型在终端设备上的应用潜力进一步被挖掘，为用户带来更加安全、高效、便捷的智能体验。未来，随着技术的不断演进，类似 OpenELM 的小模型或将在消费电子、医疗健康、工业物联网等领域发挥更大的作用。

(二) 国内小模型的应用案例

1. 垂直领域小模型智海·三乐：智能教学

智海·三乐（wisdomBot）是由浙江大学等单位设计研发的一款面向专业领域教育模型。该模型以通义千问 Qwen-7B 为技术基座，基于核心教材、领域论文和学位论文等教科书级高质量语料和专业指令数据集继续预训练和微调，集成了搜索引擎、计算引擎和本地知识库等功能。目前，智海-三乐已通过阿里云灵积平台（DashScope）对外提供 API 服务。

智海·三乐是教学一体的教育模型，既能作为老师的 AI 助教、也能作为学生的学习助手。如图 4 所示，对于老师而言，它具备教学案例生成、习题设计、作业批改、教学评估等功能。对于学生来说，它则提供知识问答、教学解析和学习建议等多种服务。



图 4 智海·三乐功能图

资料来源：wisdomBot

具体来看，图 5 展示了模型在教育领域的三大应用场景：知识问答、试题生成和智能导学，体现了其多功能性和智能化特点。以支持向量机为例，首先，学生通过知识问答来获取关于支持向量机（SVM）的知识。模型能够精准解答“什么是支持向量机”“其工作流程”等问题，并清晰地讲解关键步骤，如数据准备、训练和优化等。其次，模型可根据需求生成测试题，并提供正确答案及干扰选项，以评估学生的学习效果。最后，模型还具备智能导学功能，根据学生的基础情况，提供个性化的学习建议，涵盖基础概念、理论知识及实践操作，帮助学生制定清晰的学习路径。



图 5 智海·三乐应用场景

资料来源: wisdomBot

使用 C-Eval⁴对其进行教学评测(见表 10), 结果显示, 与通义千文 7B 相比, 经过训练后的智海·三乐在各项教育相关任务中的表现得到了显著提升, 展现出卓越的教学辅助能力。

表 10 教学评测

类别	Qwen-7B-Chat	wisdomBot
计算机网络	42.10	68.42
操作系统	42.10	47.37
计算机体系结构	42.85	47.62

⁴ C-Eval 是一个全面的中文基础模型评估套件。它包含了 13948 个多项选择题, 涵盖了 52 个不同的学科和四个难度级别。

大学编程	37.83	51.35
概率与统计	22.22	44.44
离散数学	0.00	18.75
人工智能	54.01	57.66
平均分	34.44	47.94

资料来源：智海·三乐

2. 垂直领域小模型夫子·明察：法律咨询

夫子·明察司法是由山东大学等单位联合研发的一款中文司法模型，旨在为用户提供全方位的法律咨询与解答服务。该模型以 ChatGLM-6B 为模型底座，基于海量中文无监督司法语料（包括各类判决文书、法律法规等）与有监督司法微调数据（包括法律问答、类案检索）进行训练。

夫子·明察支持法条检索、案例分析、三段论推理判决以及司法对话等功能（见表 11）。一是法条检索回复功能。对于用户的咨询，夫子·明察大模型基于生成式检索范式先初步引用相关法条，再检索外部知识库对所引法条进行校验与确认，最终结合这些法条进行问题分析与回复生成。这保证生成的回复能够基于与问题相关的法律依据，并根据这些依据提供深入的分析和建议。

二是基于案例检索回复。夫子·明察能够根据输入的案情，

分析与之相似的历史案例，并生成相似的案情描述和判决结果。生成的判决依据相关案例的法律条文，使得分析更加合理和有依据。用户可以对照这些相似案例，从中更清晰地了解潜在的法律风险。

三是三段论推理判决。司法三段论是把三段论的逻辑推理应用于司法实践的一种思维方式和方法。针对具体案件，夫子·明察能够自动分析案情，识别关键的事实和法律法规，生成一个逻辑严谨的三段论式判决预测，从而帮助用户更好地理解案件的法律依据和潜在风险。

表 11 夫子·明察三大特色效果展示

基于法条检索回复	
问：假冒他人专利进行买卖算不算诈骗？	夫子明察：根据刑法第二百一十六条规定，假冒他人专利，情节严重的，处三年以下有期徒刑或者拘役，并处或者单处罚金。因此，如果假冒他人专利进行买卖，就属于刑法所规定的“诈骗”行为，应该受到法律的惩罚。
问：公司为了发展增加设备，需要增加资金，变更股东所持比例的增资文件怎么写？变更手续需要哪些些？	夫子明察：根据公司法第一百七十八条规定，股东认缴新增资本的出资，应当以书面形式订立认缴协议，并按照公司章程规定缴纳出资。因此，变更股东所持比例的增资文件应该包括：1.变更股东名称和住所；2.变更后股东所持股份比例；3.变更后股东认缴新增资本的出资方式和期限.....
基于案例检索回复	
问：被告人蔡吉同在 2017 年 8 月 20 日下午，无证驾驶一辆无牌五征农用汽车在淮阳县冯塘乡某路段，与前方的葛某驾驶的大阳电动三轮车	夫子明察：根据上述事实，被告人蔡吉同无证驾驶无牌农用汽车，违反交通法规，导致交通事故，致一人重伤，其行为已构成交通肇事罪。根据被告人的犯罪事实、情节和悔罪表现，依照《中华人民共和国刑法》第一百三十三条、第六十七条

<p>相撞。事故导致葛某的三轮车再撞上了路边的范某及其三轮摩托车，造成葛某和范某受伤。经鉴定，范某的伤势为重伤二级。事故责任完全在蔡吉同。请问根据上述事实，该如何判罚蔡吉同？</p>	<p>第一款及《最高人民法院关于审理交通肇事刑事案件具体应用法律若干问题的解释》第二条第二款第(二)项之规定,判决如下: 被告人蔡吉同犯交通肇事罪,判处有期徒刑一年二个月。</p>
<p>三段论推理判决</p>	
<p>问:邓州市人民检察院指控:2015年6月份以来,被告人王兆龙在其经营的邓州市穰东镇乐酷动漫城内设连线机五台(每台一个操纵杆)、捕鱼机三台(其中八个操纵杆捕鱼机两台、六个操纵杆捕鱼机一台)通过“上分退币退分”的方式,吸引多人,多次进行赌博。上述事实,公诉机关提供了相关证据,认为被告人王兆龙为他人赌博提供场所、提供赌具,其行为已构成。</p>	<p>夫子明察:</p> <p>犯罪要件:</p> <p>犯罪主体: 犯罪主体为被告人王兆龙,构成要件为完全行为能力人。</p> <p>犯罪客体: 犯罪客体为社会公共秩序,即被告人王兆龙在其经营的动漫城内提供赌博场所和赌具,扰乱了社会公共秩序。</p> <p>犯罪主观要件: 《中华人民共和国刑法》第三百零三条【开设赌场罪】,因在动漫城内提供赌博场所和赌具,扰乱了社会公共秩序,被告人王兆龙构成开设赌场罪。刑期:三年以下有期徒刑、拘役或者管制,并处罚金。原因:被告人提供赌博场所和赌具,吸引多人进行赌博,构成开设赌场罪。</p> <p>判决结果: 罪名:开设赌场罪刑期:三年以下有期徒刑、拘役或者管制,并处罚金。</p>

资料来源：夫子·明察

图图 6 展示的是夫子·明察与各类法律专业模型司法能力评估体系 LawBench⁵中的表现。结果显示，夫子·明察在各类模型

⁵ LawBench 是由上海 AI 实验室联合南京大学推出的大语言模型司法能力评估体系。

中表现出色，取得了第一名，与未经法律专业知识训练的 ChatGLM 相比有了较大提升。

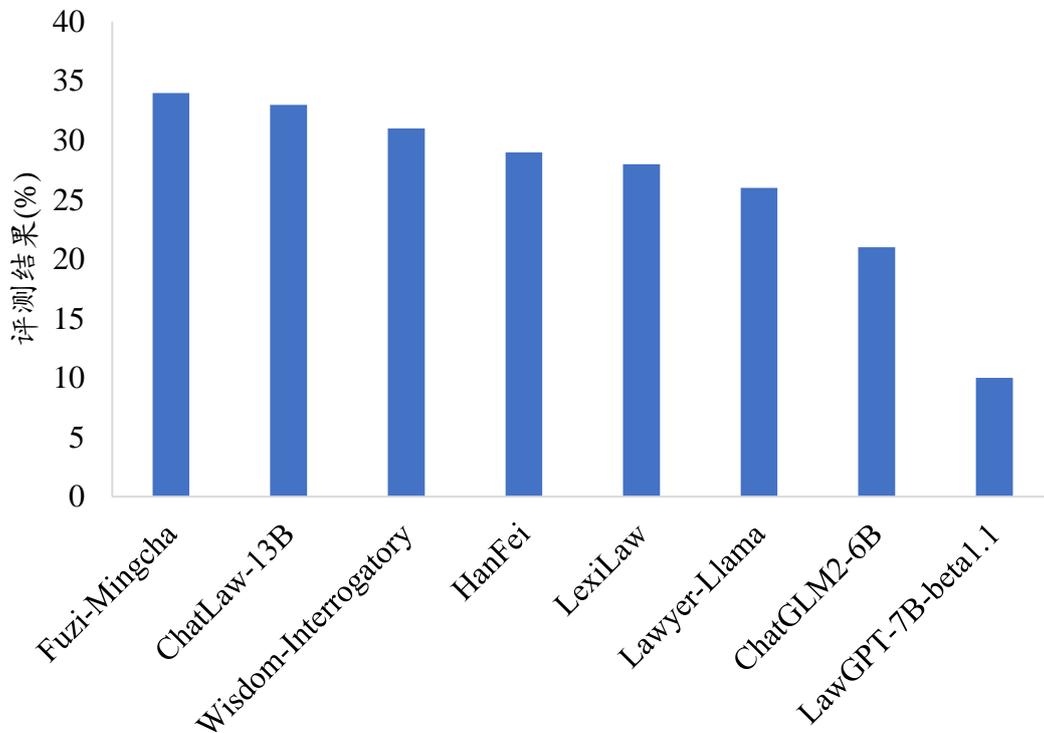


图 6 法律专业模型评测

资料来源：夫子·明察

3. 垂直领域小模型 DeepSeek-Coder-V2-Lite：代码智能

DeepSeek-Coder-V2-Lite 是 DeepSeek 推出的一款专门、用于编程语言处理、代码生成任务的小语言模型，参数规模为 16B。首先，DeepSeek-Coder-V2-Lite 支持广泛的编程语言，它包含的编程语言从 86 种增加至 338 种，几乎囊括所有主流编程语言。其次，它能够处理更长、更复杂的代码片段。它的上下文长度从

16k 扩展到 128k，能够更好地理解项目结构和上下文关系。此外，该模型完全开源，科研人员可深入研究模型架构和训练方法，开发者可自由使用和定制模型，企业可基于 DeepSeek-Coder-V2 打造专属代码智能应用。

在功能方面，DeepSeek-Coder-V2-Lite 展现出了卓越的代码生成性能。如图 7 所示，DeepSeek-Coder-V2-Lite 能够根据自然语言描述生成高质量的代码。例如，当要求“编写一个天气预报代码”时，它能够迅速提供准确且高效的实现。此外，DeepSeek-Coder-V2-Lite 还具备自动检测和修复代码错误的功能。如图 8 所示，凭借智能分析和上下文理解，系统不仅能够准确识别“天气预报”代码中的问题，还能提供详细的错误报告及修复建议。

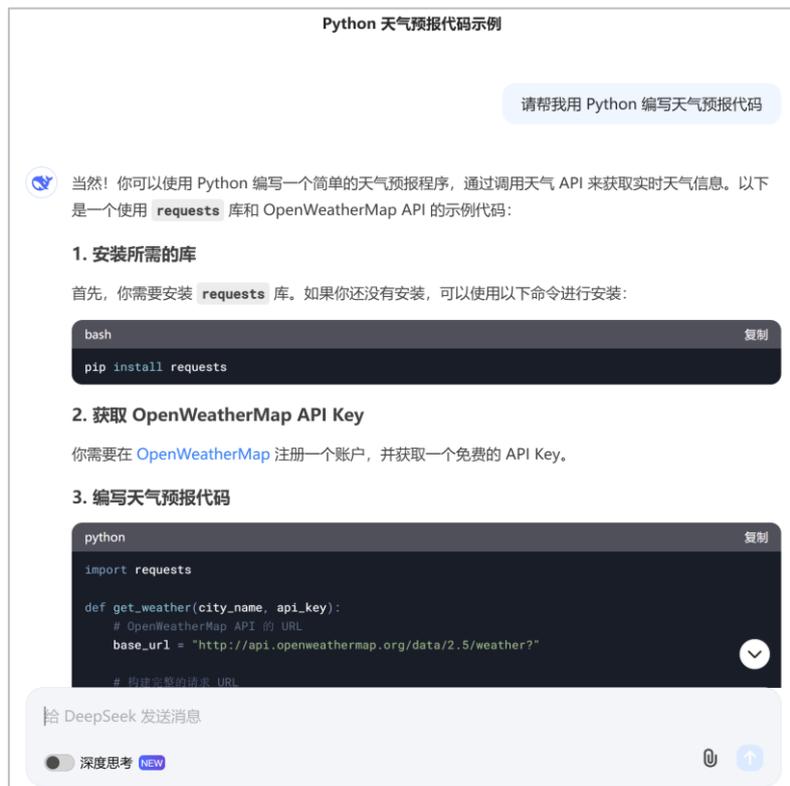


图 7 代码生成



图 8 代码修复

在编程基准评估中，DeepSeek-Coder-V2 表现优异。图 9 是 DeepSeek-Coder-V2 与 GPT-4 等通用大模型在编程基准评测中的比较。结果显示，在 HumanEval 基准测试中，DeepSeek-Coder-V2 取得了 90.2 的高分，超越了所有竞争对手。而在 MBPP+和 Aider 基准测试中也超越了 GPT4-Turbo、Claude-3-Opus 和 Gemini-1.5-Pro 等闭源模型。

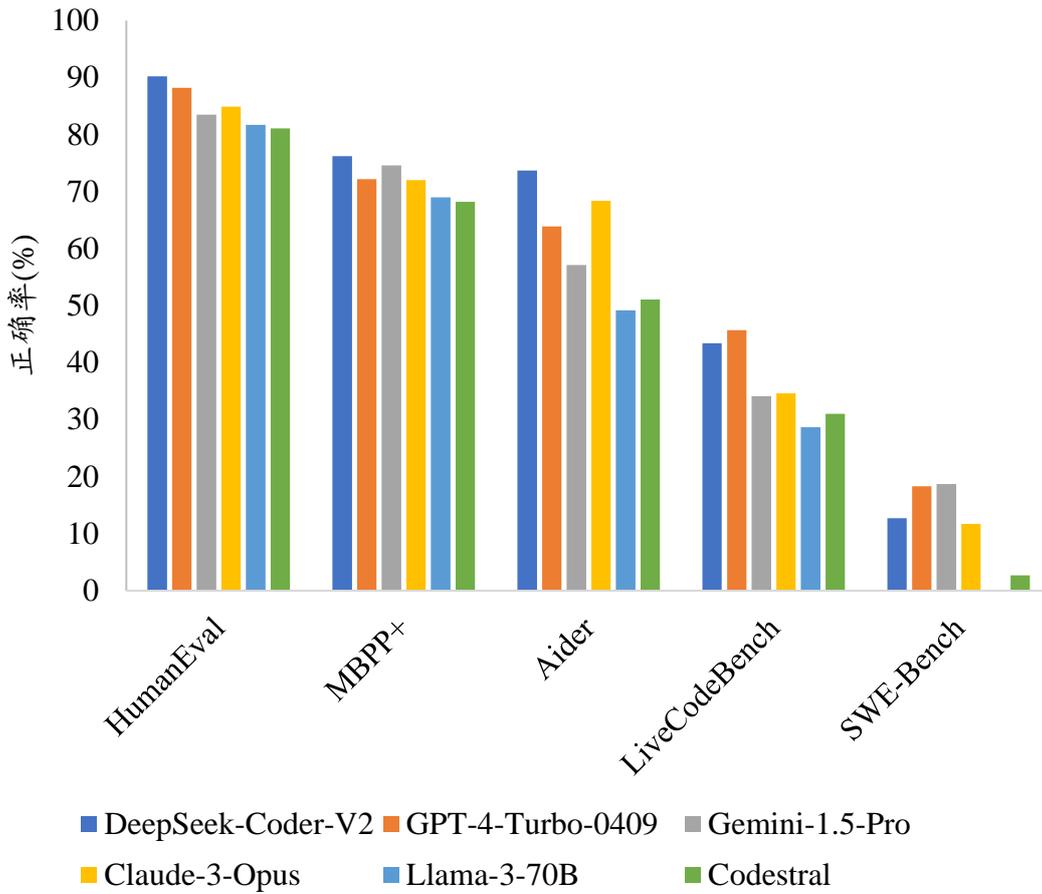


图 9 编程基准评测

数据来源：DeepSeek-Coder-V2

4. 垂直领域小模型 XuanYuan-6B：金融咨询

XuanYuan-6B 是度小满轩辕系列发布的一款金融小模型。XuanYuan-6B 模型尺寸较小，包含量化版本，降低了对硬件的要求，具备更广泛的适用性。XuanYuan-6B 结合 Self-QA 方法构建高质量问答数据，采用混合训练方式进行监督微调，使其能够有效地应对各种复杂的问答和推理任务。

XuanYuan-6B 在金融行业中具有广泛应用潜力，不仅提升了金融业务的自动化水平，还通过智能化服务优化了用户体验，帮助金融机构提高效率、降低成本、增强决策支持能力，表 12 展示的是 XuanYuan-6B 的一些金融应用场景。

表 12 XuanYuan-6B 金融应用场景

应用场景	举例
金融知识理解	金融术语解释、金融知识问答、金融法律法规、金融产品信息
金融信息摘要	金融信息抽取、金融摘要生成
金融业务分析	用户信息分析、行业趋势分析、业务数据洞察、财务报表分析
金融内容创作	金融营销文案、金融报告生成
金融计算	信贷计算、理财计算、保险计算
金融客服对话	客户意图理解、Promat 遵循、客服话术生成

资料来源：度小满

具体来看，如图 10 所示，在金融客服领域，XuanYuan-6B 能够进行智能对话，理解客户的意图，遵循 Promat（专业性对话规范），并生成符合金融行业标准的客服话术。通过自然语言理解与生成，模型能够为客户提供个性化的服务，提升客户体验。

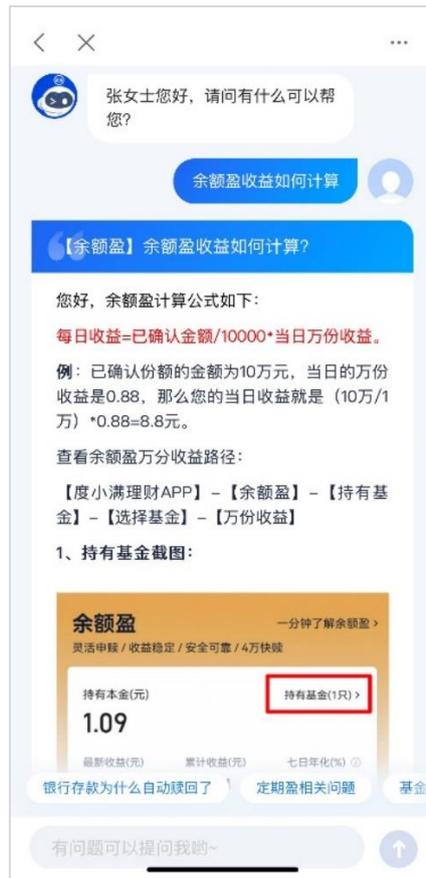


图 10 度小满金融客服对话

资料来源: 度小满

图 11 是对 XuanYuan-6B-Chat 模型进行人工评估, 来评测它在对话场景中的真实能力。评估对比对象为业界开源的类似尺寸的主流大模型, 采用 GSB (Good, Same, Bad) 指标来展示评估结果。结果表明, XuanYuan-6B-Chat 在通用性和金融能力方面的表现均优于对比对象, 显示出更强的模型能力。

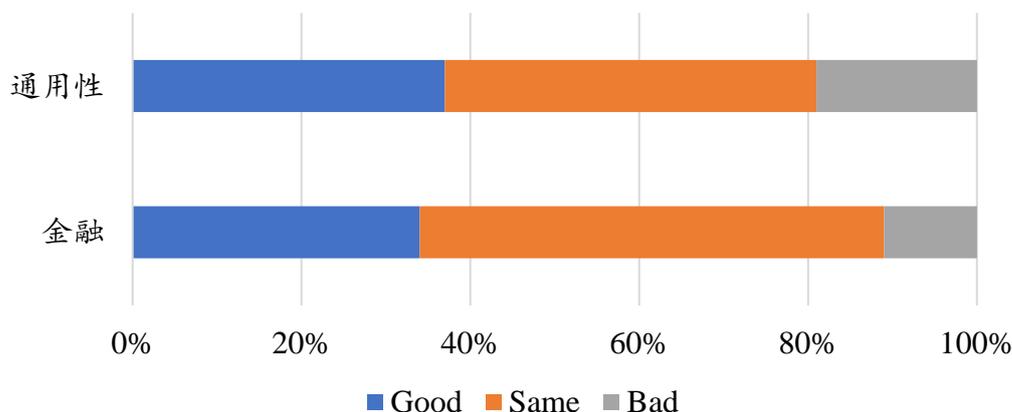


图 11 XuanYuan-6B-Chat vs 其他

数据来源：XuanYuan-6B

(三) 中国小模型的优势

中国具有丰富 AI 模型的落地场景，以及庞大的用户数据基础。在全球 AI 竞赛中，发展小模型能够帮助中国有效规避高端芯片和算力资源的制约，通过以灵活、高效、经济的方式，在特定领域快速形成技术优势并实现弯道超车。

1. 在专业领域中的表现超越国外通用模型

我国小模型体积虽小，但对具体场景中的优化能力能够超越国外通用模型。如表 13 所示，以度小满的 XuanYuan-6B 为例，将模型在金融自动评测集 FinanceIQ 测试集上进行测试，FinanceIQ 涵盖了注册会计师（CPA）、税务师、经济师、银行从业资格、基金从业资格、证券从业资格、期货从业资格、保险从

业资格（CICE）、理财规划师等十大权威金融领域考试。结果显示，XuanYuan-6B 在该项评测任务上超过 GPT-4 的水平，在知识层面展示出金融领域专家的水平。

表 13 金融评测

类别	GPT4	XuanYuan-6B
平均分	60.05	68.87
注册会计师	52.33	71.62
银行从业资格	68.72	77.65
证券从业资格	64.8	74.83
基金从业资格	68.81	77.52
保险从业资格	68.68	73.56
经济师	75.58	86.54
税务师	46.93	59.02
期货从业资格	63.51	70.67
理财规划师	63.84	71.19
精算师	27.27	26.14

资料来源：Du Xiaoman-XuanYuan

2. 综合能力大幅提升

我国小模型综合能力大幅提升，在成本和效率上有明显优势。

表 14 将 Xuyuan-6B 与 LLaMA2 系列模型的通用能力进行比较。数据显示，Xuyuan-6B 不仅具备很强的金融能力，其通用能力在知识、逻辑、代码等通用能力上的表现优异，甚至可比肩更大尺寸的模型，评测结果。

表 14 Xuyuan-6B 通用能力比较

类别	LLaMA2-70B	LLaMA2-13B	LLaMA2-7B	XuanYuan-6B
中文知识推理	50.1	41.4	32.5	81.2
中文多任务语言理解	53.6	38.4	31.8	75.8
多任务语言理解	69.8	55	46.8	64.5
小学数学	54.4	29.6	16.7	32.9
代码生成	23.7	18.9	12.8	22.6

资料来源：Xuyuan-6B

MiniCPM3.0 是北京面壁智能科技有限责任公司推出的一款 4B 参数规模为的小语言模型，它提供了更全面的功能，整体能力得到了显著提升。MiniCPM3.0 体积虽小，但通用能力表现卓越。表 15 是 MiniCPM3.0 与 Gemma2 在通用能力的评测比较。结果显示 MiniCPM3.0 在中英文文本处理能力和数学能力方面表现优异，且 MiniCPM3-4B 理论上支持处理无限上下文长度，其基准测试表现可与 GPT-3.5-Turbo 和 Phi-3.5-mini 相媲美。

表 15 通用评测

模型	MMLU (英语)	CMMLU (中文)	MATH (数学)	HumanEval+ (代码)	BFCL (工具使用)	平均分
Gemma2-9B	72.6	59.5	46.0	61.6	19.2	51.8
Llama3.1-8B	69.4	55.8	51.9	62.8	73.3	62.6
GPT-3.5-Turbo	69.2	54.5	41.8	66.5	75.4	60.2
Phi-3.5-mini (3.8B)	68.4	46.9	46.4	68.9	48.4	55.8
MiniCPM3.0-4B	67.2	73.3	46.6	68.3	76.0	66.3

资料来源：MiniCPM

除此之外，MiniCPM 具备多语言和多模态能力，能够服务不同语言的更大群体。如表 16 所示，MiniCPM-Llama3-V 2.5 采用基于 VisCPM（Visual Cross-Modal Pretraining Model）⁶的多语言训练方法，从而有效提高多语言处理能力。结果表明，在使用不到 0.5%的多语言 SFT（Supervised Fine-Tuning）⁷数据时，该模型在所有语言的性能均提升了 25 分以上。

表 16 多语言训练的效果

方法	法国	德国	葡萄	西班	捷克	匈 牙	日本	韩国	泰国

⁶ VisCPM（Visual Cross-Modal Pretraining Model）是一种多模态语言模型，旨在通过跨模态预训练方法，增强模型在视觉与语言任务中的能力。

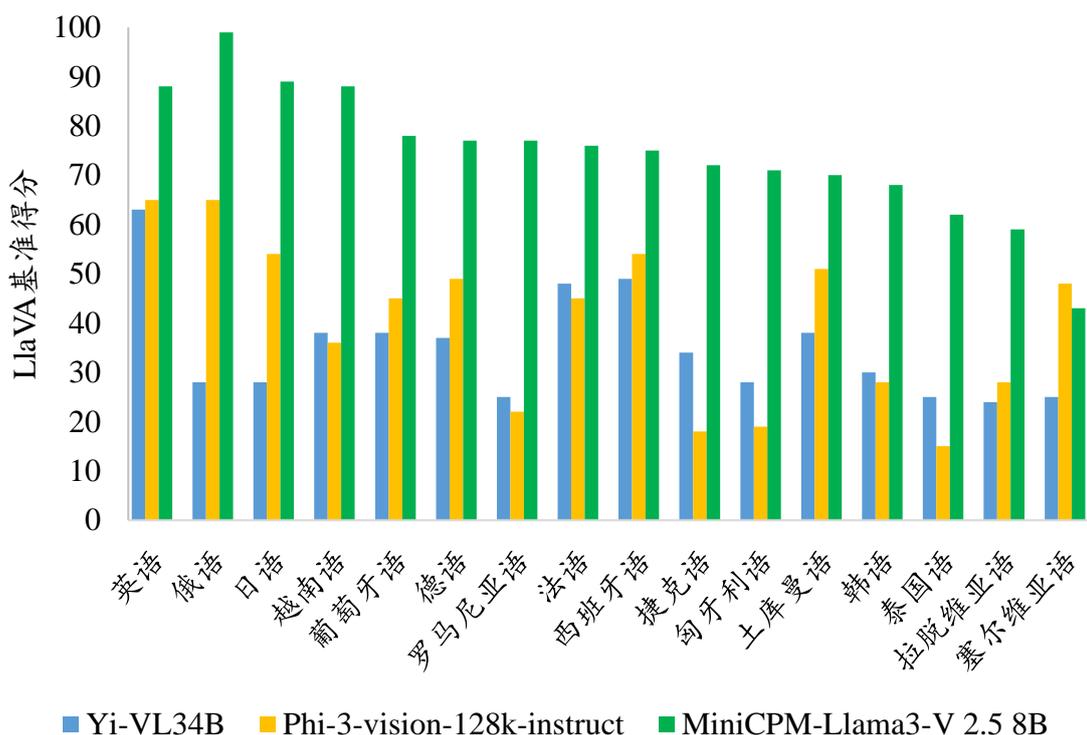
⁷ SFT（Supervised Fine-Tuning）是指在已经训练好的预训练模型基础上，使用标注数据进行进一步的微调，以适应特定任务或领域的技术。

			牙	牙		利			
w/o ML	46.4	22.8	53.0	29.0	26.5	20.6	13.8	13.7	14.4
w	72.7	76.5	83.8	73.9	71.6	70.9	88.0	67.9	61.9

(使用 90k 多语言数据 (< 0.5% SFT 数据) 进行 SFT 后训练, 并显示性能变化, ML: 多语言训练)

资料来源: MiniCPM-V Team

基于 VisCPM 的多语言训练方法, MiniCPM-Llama3-V 2.5 将其多模态能力扩展到 30 多种语言。如图 12 所示, MiniCPM-Llama3-V 2.5 可以在多语言基准测试 LLaVA Bench⁸ 上胜过 Yi-VL 34B 和 Phi-3-vision-128k-instruct。



⁸ LLaVA Bench (LLaVA Benchmarks) 是一个用于评估多模态语言模型性能的基准测试框架, 尤其是针对视觉-语言理解任务, 如视觉推理、图像-文本匹配等。

图 12 多语言多模态交互实验结果

资料来源：MiniCPM-V Team

我国小语言模型在通用能力上取得了显著提升，同时在多语言支持方面展现了巨大潜力。尽管参数量大幅减少，但小模型仍能达到与大型模型相当的性能水平，且能够有效服务于不同语言的用户。

3. 具有技术创新性

我国小模型具有技术创新性。以 **DeepSeek-Coder-V2** 为例，其参数规模为 **16B**，激活参数仅为 **2.4B**，被专门设计用于编程语言处理、代码生成和数学推理任务。**DeepSeek-Coder-V2** 采用了 **Mixture-of-Experts (MoE)** 框架，这种创新设计使得模型能够根据输入的特征（如代码片段、数学公式或自然语言描述）动态选择最合适的专家进行处理，而不是每次都启用所有专家。这种灵活的专家选择机制有效节省了计算资源，提高了计算效率与响应速度，特别适用于资源有限的应用环境。值得注意的是，**DeepSeek-Coder-V2** 的上下文窗口扩展至传统模型的 8 倍，从 **16K** 标记提升至 **128K** 标记，使得模型能够处理更长的输入，捕获更广泛的上下文信息。此外，**DeepSeek-Coder-V2** 支持多达 **338** 种编程语言，具备理解与生成多种编程语言代码的能力，适应不同

的编程环境。其训练数据集的多样性也为其优异表现提供了保障，数据集中包括 60%的源代码、10%的数学语料库以及 30%的自然语言语料库。

通过采用 Mixture-of-Experts 框架，并结合大规模参数量、扩展的上下文窗口和丰富的训练数据集等技术创新，DeepSeek-Coder-V2 在代码生成、编程语言处理以及数学推理等基准测试中超越了国外 100B 参数的 GPT4-Turbo 等通用大模型(见表 17)，展现了我国小语言模型的技术创新力。

表 17 DeepSeek-Coder-V2 表现对比

模型	开源	代码生成				数学算数		
		MBPP Plus	Human Eval	LiveCode Bench	Aider	MATH	GSM8K	2024 高考数学卷 1
GPT-4o-0513	否	73.5	91.0	43.5	72.9	76.6	95.8	52.0
DeepSeek-Coder-V2	是	76.2	90.2	43.4	73.7	75.7	94.9	60.0
GPT-4-Turbo-0409	否	72.2	88.2	45.7	63.9	73.4	93.7	47.0
Claude-3-Opus	否	72.0	84.8	34.6	68.4	60.1	95.0	46.0
Gemini-1.5-Pro	否	74.6	83.5	34.1	57.1	67.7	90.8	50.0
GPT-4-1106-preview	否	72.8	82.2	40.1	65.4	64.1	91.4	31.0
Y-Large(零一万物)	否	58.7	82.3	31.0	49.6	62.4	92.6	21.0
Qwen2-72B(阿里)	是	63.1	86.0	28.3	52.6	59.7	91.1	63.0

Baichuan4(百否川)	58.5	74.4	21.2	36.8	47.0	90.5	29.0
----------------	------	------	------	------	------	------	------

资料来源：DeepSeek

四、小模型面临的挑战及解决方案

(一) 处理复杂任务能力有限

小语言模型由于参数较少，往往在处理需要深入理解上下文、生成流畅多样的文本或应对复杂语言模式的任务时表现出较大局限性。具体来说，SLM 的能力受限于其较小的网络结构，这使得它难以有效地捕捉长篇文本中的深层次语义关系，从而导致生成的文本可能缺乏连贯性或无法准确表达复杂概念。此外，由于训练数据的局限，SLM 可能在面对非典型或少见的语言模式时出现误解或偏差。

为了提高小模型完成复杂任务的能力，一是适当增加模型的参数量，提高其学习能力和表达能力，使其能更好地捕捉复杂的语言模式和深层语义。二是提升数据集的多样性和质量，确保包含更丰富的语言样本，帮助模型更好地应对专业任务。三是通过迁移学习的方式，将 SLM 应用于特定任务时，先通过大模型进行预训练，再在特定任务上进行微调，从而提高其针对性和精度。

(二) 数据质量依赖性

小语言模型的性能高度依赖于其训练数据的质量。较小的训练数据集通常不足以覆盖任务的多样性或复杂性，容易受到数据中的偏差影响。如果训练数据中存在偏差或不平衡，SLM 可能会在特定场景下表现出较低的泛化能力，甚至产生不准确的预测或生成内容。

因此，要确保训练数据的质量。一是在数据清洗时要进行去噪处理，通过去除重复项、纠正标注错误等方式来提高数据质量。二是数据补充，结合外部的知识库，弥补数据中存在的不足，增强模型的泛化能力。此外，通过知识蒸馏 (Knowledge Distillation) 技术，将大模型的知识迁移到小模型上，提高其能力。

(三) 使用局限性

小语言模型通常是特定领域或任务设计的，这些模型的知识库相对有限，无法应对跨领域的复杂问题，因此它们在处理多领域任务时存在一定的局限性。

为了解决小模型使用局限性的问题，一是模块化设计，将小语言模型设计为可扩展的模块系统，可以通过引入专门的插件或外部知识库，扩充其对不同主题的处理能力。二是多模型集成，通过将多个小语言模型进行集成，形成一个组合型的系统，充分

发挥每个模型在特定领域的优势。

五、小模型的未来发展趋势和建议

(一) 发展小模型有助于提升我国在全球科技竞争中影响力

在全球 AI 科技竞赛中，探索合适的发展路径对于超越对手至关重要。大模型的发展对算力需求激增，然而高端 AI 芯片是关键。美国以大模型为主导的 AI 发展模式成本高昂，加之其对中国技术出口的限制，包括高端芯片的制裁，进一步加剧了中国在大模型领域的技术瓶颈。

面对这些挑战，中国需要另辟蹊径，发展小而精的模型可以有效规避资源制约，快速实现技术突破。这一发展路径更加契合中国“以创新驱动发展”的战略需求。相比耗资巨大的大模型，中国的人工智能产业链已较为完善，创新型企业与高校科研机构众多，为小模型的研发和应用提供了坚实支撑。同时，中国拥有丰富的产业生态与实际应用需求，能够更高效地推动技术普及和产业转型。

(二) 中国发展小模型的建议

一是精准优化与数据赋能打造高效小模型。未来，语言模型的发展将不再仅依赖参数规模的扩张，而是更加注重模型结构优

化与高质量数据的利用。通过优化提示工程(Prompt Engineering)和高效微调(Fine-Tuning)技术,小模型能够显著提升模型表现力。同时,利用高质量、跨领域的多模态数据,将极大增强小模型的泛化能力与应用场景的多样性,推动其在垂直领域和本地化应用中的广泛落地。

二是科技创新与生态建设推动小模型多元化发展。小模型的多元化发展需要依托科技创新和生态建设的双轮驱动。专用 AI 芯片和高效算法的研发将显著降低模型能耗与经济成本,为小模型的实际应用奠定坚实的技术基础。因此,我国要坚持高水平科技自立自强,加速关键核心技术的突破和创新,实现关键核心技术自主可控,打破外部技术封锁和制裁的束缚。同时,我国应构建开放包容的创新生态系统。扩大国际科技交流与合作,积极融入全球创新网络,深度参与全球科技创新治理,努力建设更加公平合理、开放包容、互惠共享的国际化科研环境。通过吸引国际顶尖人才,为我国 AI 领域的崛起注入活力。

中国发展小模型,能够有效规避资源和技术限制,发挥我国自身优势,充分释放小模型在垂直场景中的潜力,实现人工智能领域的全面突破和产业升级,助力我国在全球 AI 竞赛中脱颖而出。

参考文献

- [1] Apple. (2024). Apple Intelligence Foundation Language Models.
- [2] Apple. (2024). OpenELM: An Efficient Language Model Family with Open Training and Inference Framework.
- [3] DeepSeek. (2024). DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence.
- [4] IBM. (2024). Granite Code Models: A Family of Open Foundation Models for Code Intelligence.
- [5] Microsoft. (2024). Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.
- [6] MiniCPM-V Team. (2024). MiniCPM-V: A GPT-4V Level MLLM on Your Phone.
- [7] Schick, T., & Schütze, H. (2020). It's not just size that matters: Small language models are also few-shot learners.
- [8] Zhu et al. (2024). MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. King Abdullah University of Science and Technology.