本科生毕业论文开题报告

学生姓名	周彤昕	**学号**	21312832
所在院系	中山大学集成电路学院		
专　　业	微电子科学与工程		
指导教师姓名	廖思宇	**指导教师职称**	预聘助理教授
指导教师单位	中山大学集成电路学院		

毕业论文题目：基于大语言模型的中文模型合并方法研究及其应用

国内外关于本选题的研究现状、水平和发展趋势，选题研究的目的和意义等

一、国内外关于本选题的研究现状、水平

（一）国外研究现状

大型语言模型（LLM）是通过大量数据训练得到的深度学习模型，能够理解和生成自然语言。近年来，随着计算能力的提升和大规模数据集的涌现，LLM 在语言理解、生成和多种应用领域（如对话系统、翻译和文本生成）上取得了显著进展。然而，从头开始训练高性能的 LLM 是一个昂贵且复杂的任务，仅计算成本就高达数亿美元。相对而言，预训练的 LLM 可以通过微调以低成本适应新任务，导致适合特定用途的模型数量激增。

模型合并广泛应用于语言和图像生成模型，能够通过减少不良输出、提升推理效率和应对多样化任务需求，增强模型的鲁棒性和多功能性。特别是对于资源受限的语言领域，模型合并显著降低了训练成本。

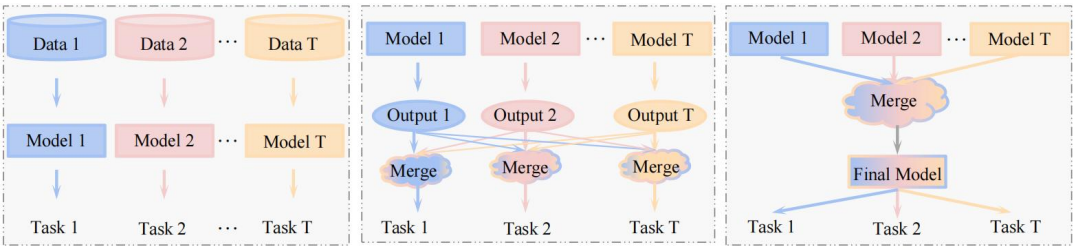

图 1 集成学习范式与模型合并范式的说明。(a) T 为 T 任务分离模型，(b)为 T 任务组装 T 分离模型，(c) A 合并模型。[2]

最近的研究表明，专门的微调模型可以迅速合并（Model Merging），以结合不同模型的能力并泛化到新的技能。模型合并是一种通过结合多个预训练模型的知识来构建新模型的方法，旨在提升模型的性能和适用性，同时保持计算效率。

模型参数合并在国际学术界已成为提升 LLM 构建效率的重要方法。近年来，该技术在各类基础模型的应用中展现了显著的优势，尤其是在生成式模型和多模态模型中，合并多个经过特定微调的模型可以实现不同能力的融合，以达到更广泛的应用效果。

现有的模型合并方法被分为预合并（pre-merging）方法和期间合并（during-merging）方法。(i)预合并方法旨在为合并创造更好的条件。它又分为使用线性化微调来实现权重空间和输入空间的解耦，进行架构转换以将异构模型转换为同构模型，以及对齐权重以将它们置于同一盆地。(ii)期间合并方法侧重于设计将多个模型合并为一个模型的复杂技术。这些方法解决了合并模型时的任务冲突和干扰问题。它们可进一步分为执行最简单参数合并策略的基本合并方法；根据特定规则计算的重要性合并多个模型的加权合并方法；将多个模型投影到稀疏子空间进行合并的子空间合并方法；在推理过程中根据输入样本动态合并模型的基于路由的方法；以及修正合并模型的基于后校正的方法。

模型合并可以应用于各种基础模型，包括大语言模型、多模态大语言模型和图像生成模型。例如，大语言模型中的模型合并有助于减少输出的不真实性，完成知识的反学习，并加快训练速度。此外，模型合并还出现在不同的机器学习子领域，如持续学习、多任务/多领域学习、少次元学习等子领域，以解决各种挑战。例如，在持续学习中，模型合并可以减轻对旧任务的灾难性遗忘。在多任务学习、多目标学习和多领域学习中，它可以促进知识转移。此外，在对抗学习中，模型合并可用于攻击和防御策略。

（二）国内研究现状

在国内，随着大语言模型（LLM）研究的迅速发展，部分基于中文处理的优秀语言模型涌现，为中文自然语言处理领域的进步奠定了坚实基础。这些模型在理解和生成自然语言方面展示了显著的能力，尤其是在中文问答和对话系统中表现突出。

尽管中文模型的发展取得了一定成就，但在模型合并技术的研究与应用方面仍处于初步阶段。现有的研究多集中于单一模型的优化，而对如何有效融合多个模型的优势进行探索尚显不足。这一领域的不足之处在于，缺乏系统化的方法论和有效的实践案例，导致模型合并的潜力未能得到充分挖掘。数据集的匮乏和缺乏统一的评测标准也限制了中文模型合并研究的深入推进。许多研究者面临着难以获取高质量、具代表性的多样化数据集的问题，从而影响了模型合并的效果和可靠性。

相比之下，国际上在此方向已有较为成熟的应用和实验，国内尚缺乏专门的中文大模型合并方法，尤其是在医疗、教育等特定领域中模型集成和迁移学习的研究还较为薄弱。

然而，模型合并的广阔前景为该领域的发展提供了新的机遇。通过将 ChatGLM、QWen 等优秀模型进行合并，可以实现不同能力的融合，提升模型的适用性与性能。这种合并不仅可以减少从头训练新模型所需的高昂成本，还能够更好地应对多样化的应用需求，从而加速中文大语言模型的创新与推广。

因此，研究中文模型合并技术具有重要的意义，不仅能够推动相关领域的学术发展，还能为实际应用提供更强大的技术支持。随着对模型合并技术的深入研究，未来将有望在多领域、多任务的应用中取得突破性进展。

在中文大模型的合并与迁移学习研究中，以下表现突出的语言模型具有重要参考价

值：

ChatGLM 是一个基于 General Language Model 架构的支持中英双语的开源对话语言模型，具有 62 亿参数。结合模型量化技术，用户可以在消费级显卡上进行本地部署。由于 ChatGLM-6B 优越的中文处理性能与开放的权重参数，其具有优越的参考意义。

Qwen 大模型覆盖多语言，以中文和英文为主，其系列工作均被开源，具备稳定且大规模的数据。在相关基准评测中，Qwen 系列模型体现出了非常有竞争力的表现，显著超出同规模模型并紧追一系列最强的闭源模型。此模型也将作为本项目的重点研究参考。

ShenNong-TCM 模型以 LlaMA 为底座，采用 LoRA(rank=16)微调得到。其训练数据为中医药指令数据集，以开源的中医药知识图谱为基础，采用以实体为中心的自指令方法（entity-centric self-instruct），调用 ChatGPT 得到 11 万以上的围绕中医药的指令数据。其在中医药领域的数据训练，适用于中文大语言模型在特定领域的的迁移与合并。

视觉-语言-动作模型（VLA）是一种多模态模型，旨在处理视觉和语言信息，并输出相应的机器人动作。VLA 能够将视觉信息（如图像或视频）与自然语言输入相结合，以理解场景并生成相应的动作指令。利用大语言模型的语言理解和生成能力可以增强 VLA 在多模态人物中的表现，进而提升多模态输入处理、自然语言指令的生成、人机交互、多任务学习等能力。

二、关于本选题的研究水平

目前，模型合并技术在国际范围内已取得显著进展，并展现出较高的技术成熟度。诸多方法如参数平均、专家合并、模型堆叠等，已被广泛应用于不同类型的模型，尤其是在多模态和生成式模型中。

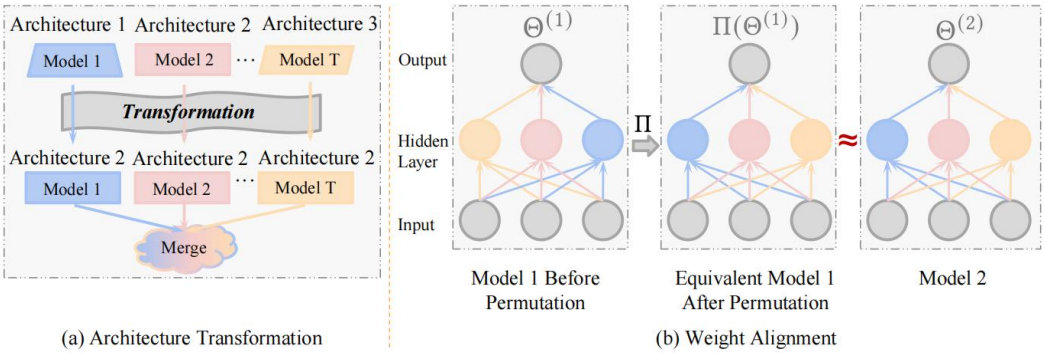

图 2：（a）是一个架构转换的说明，它将多个异构模型转换为同构模型，允许执行后续的直接参数级合并操作。（b）一个权值/参数对齐的说明，即排列神经网络模型 Θ（1），使其与模型 Θ（2）对齐。[2]

在近期的 NeurIPS 社区中，提出了“LLM 合并挑战”[1]，探讨了是否可以通过合并现有模型而非从头开始微调或训练另一个 LLM，以满足新的技能和设计参数的需求。该比赛侧重于利用最少的计算和额外参数，合并 Hugging Face 公开发布的现有专家模型，目标是开发出超越现有模型和合并基线的合并模型。

大型语言模型社区已开发出多种模型合并的方法：

· **参数平均模型合并。**包括简单平均、Fisher 合并、任务算术、TIES-Merging 等。此外，参数平均合并方法还常用于合并在不同风格上单独微调的 Stable DiffusionXL 模型，以生成结合这些风格的图像。

· **MoE（专家混合）模型合并。**在 MoE 模型中合并专家以加速推理。将模型合并为新的 MoE 模型，通过基于 token 的门控机制。学习后置路由器以在专业化 PEFT 模块之间进行选择。

· **模型堆叠方法。**将多个 LLM 按顺序应用或组合其输出。自校正器（self-correctors）和对齐器（aligners）将一个 LLM 的生成结果作为输入传递给另一个 LLM 以产生最终输出。代理微调（proxy-tuning）结合来自较大基础模型和较小微调模型的下一个 token 预测。控制解码（controlled decoding）通过前缀评分的 LLM 调整下一个 token 概率以使其与奖励函数对齐。

· **模型压缩（Model Zipping）。**通过交叉注意力结合不同 LLM 的层，例如 CALM 和 Frankenllama，通过扩展隐藏维度进行合并。

· **模型路由方法。**旨在通过较小的路由模型选择最合适的 LLM 进行每个输入，同时保持候选 LLM 不变。例如，LLM Blender 训练一个小模型对候选生成结果进行排序，FoE 利用候选专家 LLM 的 token 嵌入来选择最佳模型，其他方法训练元模型预测某个 LLM 是否会为输入产生良好的生成结果。

尽管中文模型在性能上取得了一定成就，但与国际先进水平相比，仍存在技术差距。当前，中文大语言模型在合并技术的应用上相对滞后，尚未形成系统化的方法论和丰富的实践案例。此外，针对中文模型的基准测试和评估标准也缺乏一致性，这导致国内研究在验证和比较不同模型合并效果时面临困难。

在具体应用方面，已有研究表明，合并技术在资源受限环境中具有显著的优势，能够有效降低训练成本，并提升模型的鲁棒性和泛化能力。然而，中文模型合并的实践应用仍处于探索阶段，尚未充分利用现有的高性能模型进行组合。

综上，中文模型合并技术的发展潜力巨大。随着技术的不断成熟和方法的逐步完善，未来有望在各个领域实现更广泛的应用，并在实际效果上超越当前的技术水平。这为推动中文自然语言处理技术的进步提供了良好契机。

三、关于本选题的发展趋势分析[2]

（一）模型合并技术的发展趋势分析

虽然目前有大量先进的模型合并方法和应用，但现有的模型合并方法的技术和应用仍然面临挑战。未来的发展趋势可能将聚焦以下方向。

1. 缩小合并模型与独立模型之间的性能差距。

在实际应用中，保证模型合并的性能仍然具有挑战性。当前模型合并技术的有效性

在很大程度上依赖于“预训练-微调”范式。具体来说，成功的模型合并需要基于同一预训练模型对多个模型进行微调，并在微调过程中对历时数和学习率进行仔细控制。如果这些超参数设置不当，模型可能无法在相同或接近的盆地中收敛。即使基于预训练的微调范式，合并后的模型与独立模型之间仍存在明显差距，尤其是当模型/任务数量较多时。因此，未来有希望的研究方向是探索如何在更宽松的条件下确保模型合并的有效性。例如，研究如何在不影响性能的情况下合并针对不同任务从头开始独立训练的多个模型可能很有价值。

2. 模型合并的深入理论分析。

现有模型合并技术的有效性和解释主要是经验性的，缺乏足够的理论保证。目前关于模型合并的理论研究或解释还很有限。现有的少数研究主要集中在合并在同一轨迹上或同一数据集上经过不同微调设置训练的多个模型。关于合并在不同数据集上微调的多个模型或合并在不同数据集上从零开始训练的多个模型，几乎没有任何理论研究或解释。因此，未来的研究应着眼于更全面、更深入的理论分析，以提高模型合并的成功率和可靠性。此外，深入了解模型合并的有效性反过来也有助于发现更有利于模型合并的先验条件。

3. 值得信赖的模型合并。

模型合并容易引发知识产权纠纷和中毒攻击，因此开发可靠可信的合并方案成为研究的当务之急。对模型合并可靠性的研究可分为两个关键角色：模型所有者和模型合并者。一方面，对于模型所有者来说，保护其模型的知识产权是首要问题。这种保护包括主动和被动两种防御策略：

(1) 主动防御：模型所有者可能希望确保其发布的模型被独立使用，而不被其他用户合并。主动防御策略的理想结果是，模型在按原计划使用时性能稳定，但如果与其他模型合并，则会完全崩溃。

(2) 被动防御：当模型所有者怀疑自己的模型被合并时，需要有一种稳健的方法来验证合并后的模型是否包含其原始模型。另一方面，对于模型合并器来说，如何在合并一组授权模型时有效防止恶意注入，如后门或中毒攻击，也是一个重要的研究方向。

4. 有效和高效的模型合并。

现有的高性能模型合并方法通常会在效率和内存方面付出巨大代价。首先，这些方法大多需要在执行过程中将所有模型加载到内存中。例如，合并 72 个微调 ViT-B/32 模型需要 200GB 以上的内存。此外，确定模型合并系数的启发式方法涉及对合并模型的重复评估，而可学习方法则依赖于额外的数据和训练。未来，开发无需训练、额外数据、GPU 或大量内存的更高效模型合并方法将大有裨益。

5. 合并异构专家模型。

现有方法主要侧重于合并同构模型。然而，在实践中，许多异构模型在各种任务中表现出色。数量有限的现有异构模型合并方法涉及利用知识提炼技术将多个异构模型转化为同构模型，然后再进行合并。因此，值得探讨的还有如何合并这些异构模型，而无

需承担与架构转换相关的高昂成本。

6. 模型合并的跨学科应用。

模型合并已被巧妙地应用于各种基础模型和机器学习子领域，以应对不同的挑战并完成有趣的任务。如何将模型合并策略从一个子领域调整到另一个子领域，是一个令人兴奋的探索方向。

（二）中文模型合并技术的发展趋势分析

随着人工智能技术的迅猛发展，尤其是大语言模型（LLM）和多模态模型的不断进步，中文模型合并技术的前景愈加广阔。以下是几个主要的发展趋势：

- **技术创新与优化：**未来，模型合并方法将不断创新，以提高效率和效果。新兴的技术，如自适应合并和动态模型选择，将使得模型合并更加灵活，能够根据具体任务和输入条件自我调整。此外，针对特定领域的个性化合并方法也将逐渐兴起，以更好地满足行业需求。
- **多模态融合：**随着视觉-语言-动作模型（VLA）等多模态模型的发展，未来的研究将更加注重不同模态间的融合。将语言模型与视觉、动作等信息相结合，不仅可以提升模型的整体性能，还能拓展应用场景，促进人机交互和智能机器人领域的进步。
- **应用领域拓展：**随着模型合并技术的不断成熟，预计将在医疗、教育、金融等多个特定领域得到广泛应用。通过合并具有不同专业知识的模型，能够实现更为精准的任务执行和决策支持。这将推动相关领域的智能化发展，提高行业整体效率。
- **标准化与可解释性：**在合并技术日益普及的背景下，建立统一的评估标准和可解释性框架将成为重要趋势。通过制定一致的基准测试和评估方法，研究者能够更好地比较不同模型合并方法的效果。同时，增强模型合并结果的可解释性，有助于提升用户对 AI 系统的信任。
- **开源与合作：**未来，开源项目和学术界的合作将推动模型合并技术的普及和发展。通过共享数据集、模型和研究成果，研究者能够更有效地进行合作，推动技术进步并加速创新。

综上所述，中文大语言模型的合并技术正在快速发展。通过不断的技术创新和多样化的应用实践，该领域将迎来更多的机遇和挑战，推动中文自然语言处理的整体进步。

四、选题研究的目的和意义

（一）研究目的

本研究旨在探索和开发适用于中文大语言模型的参数合并方法，构建具有较高效率和适应性的中文大模型，并提升其在特定任务中的表现，填补中文模型领域在模型合并方法上的空白。

（二）研究意义

· **技术创新性。**通过创新的合并技术，提升中文大模型在不同任务中的灵活性，为后续多模态、跨领域的中文大模型应用提供技术支持。

· **应用广泛性。**通过技术向特定领域，如本项目中提到的中医药领域的迁移，促进模型合并向应用领域的发展。在医疗、教育、法律等领域中，模型合并技术可使模型快速适应特定需求，减少训练时间和成本，并增强模型在中文语料中的表现。

· **推动中文自然语言处理的发展。**在中文模型合并技术的研究推动下，可以进一步缩小国内外在大语言模型技术上的差距，为中文自然语言处理的理论发展提供支持。

选题研究的计划进度及可行性论述等

一、选题研究的计划进度

为确保选题研究的顺利推进，以下将研究计划分为几个阶段，预计在 2025 年 3 月基本完成题目和毕业论文:

阶段一：文献调研与选题确认（2024 年 10 月 28 日 - 2024 年 11 月 15 日）

收集国内外关于大语言模型合并的研究文献，重点关注中文模型的相关研究。

确认选题的研究框架与主要方向，并整理出相关的研究问题与目标。

阶段二：研究设计与方法论制定（2024 年 11 月 16 日 - 2024 年 12 月 15 日）

明确研究方法与技术路线，确定合并技术的具体实现方式。

制定详细的实验方案，包括数据集选择、模型选择及合并策略等。

阶段三：数据收集与预处理（2024 年 12 月 16 日 - 2025 年 1 月 15 日）

收集所需的数据集，进行数据清洗与预处理，确保数据的质量与代表性。

对所使用的中文模型进行初步测试，确保模型的可用性。

阶段四：模型合并与实验实施（2025 年 1 月 16 日 - 2025 年 2 月 15 日）

进行模型合并实验，记录各类模型的性能指标。

进行多次实验以优化合并策略，确保研究结果的可靠性。

阶段五：数据分析与论文撰写（2025 年 2 月 16 日 - 2025 年 3 月 10 日）

对实验结果进行深入分析，提炼出研究结论。

撰写毕业论文，整理文献综述、研究方法、实验结果与结论。

阶段六：论文修改与答辩准备（2025 年 3 月 11 日 - 2025 年 3 月 25 日）

根据指导老师的反馈进行论文修改，完善论文内容。

准备答辩材料，进行模拟答辩，确保顺利通过最终答辩。

二、选题研究的可行性论述

（一）选题合理性

LLM 技术在当今的发展已经相当充分，模型合并也已经被最近的研究表明正确性。

随着人工智能技术的飞速发展，特别是大语言模型的广泛应用，中文模型合并的研究逐渐受到重视。当前，国内在这一领域的研究仍处于初步阶段，开展此研究不仅能够填补现有研究的空白，还能够为实际应用提供切实可行的解决方案。此外，结合已有的优秀模型（如 ChatGLM 和 QWen），进行有效的合并研究，不仅具有可行的技术方法，也具有显著的学术价值与应用前景。这一研究不仅契合国家对人工智能技术的重视，也符合行业发展的需求。

在“LLM 合并竞赛”中，给出了完整可靠的数据集与性能标准。本项目将参考其标准测试最终性能，以考察合并效果。实验给出了入门套件与工具，帮助我迅速理解与适应项目。现有大量的专家模型可供使用，重点参考的模型在上文已经列出。项目具有明确的数据集。具体如下:

Cosmos QA 是一个包含 35.6K 个问题的大型数据集,这些问题要求以常识为基础的阅读理解，并以多项选择题的形式出现。 它侧重于阅读人们日常叙述的字里行间，提出的问题涉及事件的可能原因或影响，需要在上下文的确切文本跨度之外进行推理。

Extreme Summarization (XSum) 数据集是一个用于评估抽象单篇文档摘要系统的数据集。 其目标是创建一个简短的、一句话的新摘要，回答 "这篇文章是关于什么的？ 该数据集由 226,711 篇新闻文章组成，并附有一句话摘要。 这些文章收集自英国广播公司（BBC）的文章（2010 年至 2017 年），涵盖多个领域（如新闻、政治、体育、天气、商业、技术、科学、健康、家庭、教育、娱乐和艺术）。 官方随机拆分的训练集、验证集和测试集分别包含 204,045 份（90%）、11,332 份（5%）和 11,334 份（5%）文档。

综上所述，项目具有明确的任务、标准、工具套件与参考数据集，故项目具有充分的可行性。

（二）个人能力

本人是一名微电子科学与工程专业的学生，具备扎实的理论基础和丰富的实践经验。通过参与多个与人工智能相关的项目与竞赛，我锻炼了自己的项目管理和技术研发能力。

在项目方面，我参与过多个与 AI 相关的技术项目，包括面向机器人视觉系统的深度估计硬件处理器设计等。此外，我还参与了基于 FPGA 的二维 5×5 拓扑结构的片上网络设计和 ResNet-50 的硬件优化项目，积累了丰富的硬件开发和深度学习模型应用经验。这些经历使我对人工智能模型有了深刻的理解和认识，此外，我具备较强的文献调研能力与数据分析能力，为后续研究奠定了坚实的基础。

（三）技术指导

此研究将得到廖思宇教授的技术指导。廖教授是硕士生导师，曾获罗格斯大学博士学术成就奖，专注于高性能神经网络模型及相关应用的研究。他的研究成果已在 MICRO、ISCA、ICML、AAAI 和 CIKM 等国际顶级会议和期刊上发表。廖教授还拥有丰富的工业经验，曾获得国际商业机器实习优秀奖，并入围高通创新奖学金决赛。他在多领域的深

厚背景和丰富的实践经验，将为本项目提供宝贵的技术支持和指导，确保研究的高质量和实用性。

毕业论文撰写提纲

摘要

该部分简要概述研究的目的、方法、主要结果和结论。应突出研究的创新点和贡献，吸引读者的兴趣。

1. 绪论

1.1 选题背景与意义

介绍大语言模型和模型合并的背景，说明研究的重要性以及在中文处理领域的应用前景。

1.2 国内外研究现状

概述国内外在大语言模型构建及模型合并技术方面的研究进展，指出存在的不足和挑战。

1.3 本文的研究结构

概述各章节的主要内容与结构，帮助读者理解论文的整体框架。

2. 文献综述

2.1 有关构建大语言模型的研究

详细介绍构建大语言模型的关键技术与方法，探讨不同模型架构的优缺点。

2.2 模型合并技术

综述现有的模型合并技术，分析其实现方法与应用场景。

2.3 基于中文语言模型的研究现状

着重介绍中文大语言模型的研究现状，重点分析其在自然语言处理中的表现。

2.4 现有研究的不足

指出当前研究中的不足之处，例如技术局限性、应用场景的缺乏等，为后续研究提供依据。

3. 研究方法

3.1 模型选择

阐述所选大语言模型（如 ChatGLM、Qwen 等）及其理由，分析其特性和适用性。

3.2 合并框架

介绍模型合并的具体框架与流程，强调所用的合并技术与算法。

3.3 评估基准分析

确定评估模型性能的标准和方法，包括实验设计和指标选择。

4. 项目实现

4.1 数据准备

描述所用数据集的来源、特征及预处理方法，确保数据的质量和代表性。

4.2 模型合并过程

详细记录模型合并的步骤，包括所用工具与技术的具体实现。

4.3 模型训练过程

讨论模型训练的设置、参数选择及训练过程中的关键决策。

5. 结果与讨论

5.1 性能评估

分析合并模型的性能表现，使用具体数据和图表展示结果。

5.2 应用探析

探讨合并模型在实际应用中的潜在效果与场景，讨论其对行业的影响。

5.3 挑战与限制

识别研究中遇到的挑战与限制因素，并讨论其对结果的影响。

6. 结论

6.1 总结

总结研究的主要发现与贡献，回顾研究问题的解决方案。

6.2 未来研究展望

提出未来研究的方向和可能性，探讨可以进一步深入的领域与问题。

6.3 结语

以简洁有力的方式结束论文，强调研究的意义与对未来的影响。

7. 参考文献

列出所有引用的文献，确保遵循特定的引用格式。

参考文献

[1] llm-merging. (2024). LLM-Merging. GitHub. https://github.com/llm-merging/LLM-Merging

[2] arXiv:2408.07666 [cs.LG] https://doi.org/10.48550/arXiv.2408.07666

[3] Zhu, W., Yue, W., & Wang, X. (2023). *ShenNong-TCM: A Traditional Chinese Medicine Large Language Model*. GitHub repository. https://github.com/michael-wzhu/ShenNong-TCM-LLM

[4] Gryphe. (n.d.). *BlockMerge_Gradient*. GitHub repository. https://github.com/Gryphe/BlockMerge_Gradient