附件一：2024 年 CCF-百度松果基金申报主题

目录

1. 深度学习基础理论和技术

本课题期望通过对深度学习基础理论、基础技术层面的研究，来指导解决当前深度学习研发和应用上的挑战性问题，侧重对实际问题更有直接指导意义的或者对未来技术发展有引领意义的基础研究。

建议研究方向：

1.1 优化算法和收敛性研究

概述：训练收敛效率优化是提升大模型训练效率和算法效果、支持大模型快速迭代的关键。本研究方向可从大模型收敛加速方法，从模型结构、低精度、学习率/批大小超参选取、优化器类型等多个角度出发，探索适合大模型快速收敛的途径，提升大模型收敛效果。

研究方法推荐：

（1）研究大学习率、大 Gobal Batch Size 下的加速收敛方法。

（2）研究低数值精度训练收敛加速方法。

（3）研究新型优化器收敛加速方法。

验收标准：在 LLaMA 等代表性大模型上，达到相同下游评估指标的前提下，训练消耗的计算卡时资源节省 30%以上。相关工作集成到飞桨框架中。

1.2 编译优化算子融合正确性理论研究

概述：深度学习编译器很重要的功能就是将用户可以识别的 Pattern 能够生成硬件相关的高效代码。对于 CUDA 硬件，有一类重要的优化方法：算子（OP）融合，又被称为 Kernel 融合。CUDA 硬件存在多类不同的内存结构，其中有一类称为“全局内存”的访存开销是很大的。如果我们将两个连续 Kernel 合并为一个 Kernel 调用，我们会减少中间变量的读写开销，因此在访存密集型的 2 个算子上，融合可以获取更高的性能。各种业界实践也都表明有效的算子融合技术可以极大的加速运行性能。但是，融合过程中保证性能的前提下保障正确性是一个难题。

研究方法推荐：

希望能够借助数学的语言和推理进行建模，并从精确的定义出发保障正确性。

（1）精确定义 Elementwise/Broadcast/Injective/Reduce 四类算子。

（2）定义算子融合变换。

（3）找到融合变换不影响性能并正确的充分条件。

验收标准：

（1）基于飞桨的 CINN 编译器，能够有效识别可以融合的算子，并且保障融合的正确性，静态 Shape 千级别子图达到成功率 95%以上。

（2）给出不影响性能的 Reduce+Broadcast+Reduce 融合条件，并生成正确 Kernel。

1.3 大模型二值化训练、压缩与推理研究

概述：低比特量化压缩将浮点数转为低比特整数进行计算和存储。在硬件的整数计算单元支持下，可显著提升模型推理效率。在当前大模型领域，产业界已普遍在业务中采用 8 比特量化压缩，学术界则聚焦 4 比特量化压缩，并有研究更低比特量化压缩的趋势。二值化（*binary quantization*）即 1-bit 量化压缩，则是低比特压缩的终极形式。通过二值化的研究，不仅可以提升推理效率，还有助于洞察生成式大模型底层原理，对模型结构设计优化、训练方法改进和硬件设计都有启发。

研究方法推荐：

（1）训练：研究二值化网络的训练方法，使其效率和效果综合优于传统模型。

（2）压缩：结合信息理论，对生成式大模型的注意力机制（Attention）、位置编码(Position Embedding)、残差连接(Residual)、归一化(Normalization)等关键结构的原理进行研究，指导设计适用于二值化的量化方法和模型结构。

（3）推理：研究二值化网络推理加速技术，在当前主流硬件上实现二值化推理加速，并给出对硬件的改进设计建议。

验收标准：

（1）设计新的二值化网络，比业界 SOTA 模型，在推理显存占用一样情况下，效果相对提升 10%以上。

（2）基于主流生成式大模型结构和主流生成式评估集上，经过二值化压缩后，模型效果相对损失在 10%以内。

（3）通过训练或压缩得到二值化模型，推理速度优于同等规模模型的 INT4 推理性能。

1.4 Transformer 核心计算加速技术研究

概述：在 Transformer 类大模型训练过程中，矩阵乘法、注意力计算等环节占比可达到端到端耗时的 60%以上，针对这些重点环节算子的优化对提升大模型训练吞吐具有重要意义。

研究方法推荐：针对大模型矩阵乘法、注意力计算等场景，开发高性能的 Kernel 实现。

验收标准：开发一套可针对不同 shape 和数据类型（含 bfloat16、float16 等）的矩阵乘法、注意力计算加速工具包，可自动产生最优的高性能 Kernel 代码，Kernel 性能提升 20%以上，并可自动编译运行。相关成果集成到飞桨框架中，并在 Llama 等开源大模型上应用。

2. 深度学习框架技术

本课题期望对深度学习框架技术进行系统、深入和前瞻的研究，解决当前深度学习框架中存在的关键难点问题，探索下一代框架的设计实现。

建议研究方向：

2.1 编译优化加速技术研究

概述：随着硬件算力发展速度远大于访存、CPU 调度、总线带宽的发展速度，编译优化是深度学习框架必备的一个优化技术，如何提供高效的通用加速方案是一个重要的研究方向。以 NV GPU 平台为例，当前主流的编译器能够将 reduce、elementwise、broadcast 等类型的算子做较好的融合和优化，但是对于当前 LLM 使用较多的 attention 子图，自动融合优化的相对较少。如何生成高效的 attention 子图，并且能够和前后的算子（如 ROPE 等）做更大的融合，是一个性能优化的难题。

研究方法推荐：例如通过提出一种模板的思路，能够对于 attention 结构进行建模，并且能够生成高性能的 CUDA kernel。

验收标准：基于飞桨的 CINN 编译器，能够生成高性能的 attention 的 kernel，性能能够达到 FlashAttention 的 95%+。

2.2 面向多芯片的通用算子融合大模型推理加速策略

概述：当前各类 AI 芯片都针对大语言模型（如 LLaMA 系列）的推理性能进行优化，由于芯片架构差异，硬件原生算子融合方案在算子融合颗粒度和融合算子的接口设计上差异较大；期望基于飞桨 BlockAttention 的高性能大模型推理方案，研究通用的算子融合策略，不仅在不少于 3 款 AI 芯片下都能获得较高的推理性能提速，且融合算子力度一致接口统一。

研究方法推荐：

（1）模型选择：PaddleNLP 中支持的开源大模型中，至少选择 13B 或以上参数的大模型 1 个。

（2）研究方法：分析不少与 3 款 AI 芯片（如海光 DCU、昇腾 NPU、昆仑 XPU、寒武纪 MLU 等）的大模型推理性能，包括细粒度小算子方案和硬件原生算子融合方案下的性能对比；设计通用的算子融合策略，并针对性为 3 款芯片开发对应的融合算子（如海光 DCU 支持 HIP 算子开发、昇腾 NPU 支持 AscendC 算子开发、寒武纪 MLU 支持 BANGC 算子开发等）；验证通用融合算子的推理性能，得出对比之前的细粒度小算子方案的提升效果，以及对比硬件原生算子融合方案的下降比例。

验收标准：至少在 3 款 AI 芯片（如海光 DCU、昇腾 NPU、昆仑 XPU、寒武纪 MLU 等）上验证通用的算子融合策略相较于细粒度的小算子方案的推理性能提升 20%+，且推理性能不低于硬件原生算子融合方案性能的 90%。

2.3 编译器硬件适配技术研究

概述：研究如何合理抽象硬件特征，并据此定义协议将其接入编译器，同时确保编译器能充分利用硬件特征生成高性能代码。

研究方法推荐：

（1） 硬件平台分析：详细研究和分析不同硬件平台（如 CPU、NV GPU、海光 DCU、昇腾 NPU、昆仑 XPU、寒武纪 MLU 等）的架构特性和并行计算模型。重点关注各硬件平台如何做内存合并、共享内存管理、SM 内或 SM 间的调度。

（2）硬件接入协议设计：基于硬件平台分析结果，设计编译器接入协议，包括编译期融合策略协议、编译期性能优化协议、编译期代码生成协议、运行时资源利用协议等。

验收标准：在多种硬件平台上进行多个常见模型验证，硬件平台可以选海光 DCU、昇腾 NPU、昆仑 XPU、寒武纪 MLU 等，常见模型可以选 ResNet、Bert、LLaMA2、Baichun2、ChatGLM、Qwen 等。通过与 NV GPU 的对应版本对比正确性，评估硬件适配方案的有效性和通用性。通过与该硬件上不开编译优化的对应版本对比性能，评估硬件适配方案的性能优化空间。

2.4 大模型高效分布式训练技术

概述：随着大模型的规模和复杂性不断增加，分布式训练也面临新的挑战。例如，不同的模型架构和数据规模对分布式训练策略的要求各不相同，高效、通用的大模型分布式训练技术重要性凸显。本研究方向建议研究高效通用的分布式训练技术，以解决大模型训练的挑战，包括但不限于研究自动并行技术、研究分布式训练性能优化技术等。

研究方法推荐：。

（1）调研全自动相关的工作，同时结合飞桨框架的特点以及飞桨现有的半自动并行架构，研究如何将半自动并行技术，扩展为全自动并行。

（2）深入分析大模型训练的性能瓶颈，从多个维度提升大模型训练性能，包括但不限于高性能算子开发、通信优化、通信和计算 overlap 等。

验收标准：

（1）基于飞桨框架，研发和实现通用自动并行训练技术，在不少于 3 个开源大模型上，能够全自动（无需用户标记）将单卡模型转为分布式训练，并且性能与手动并行或半自动并行版本持平或领先。

（2）基于飞桨框架，研发和实现大模型训练性能优化技术，在主流大模型上，相对于飞桨开源版本性能提升 10%+，且收敛性不下降。

3.大语言模型（Large Language Model）

近年来，大语言模型（Large Language Model）在自然语言处理领域取得了巨大的成功。大语言模型（LLM）是指使用大量文本数据训练的深度学习模型，可以生成自然语言文本或理解语言文本的含义。大语言模型可以处理多种自然语言任务，如写作、问答、对话等，是通向人工智能的一条重要途径。

建议研究方向：

3.1 基础模型训练

概述：面向基础模型进行模型结构、预训练技术、对齐技术、模型分析等方面的研究。

研究方法推荐：

（1）新型模型结构：研究具备超长序列处理能力的新型基础模型结构，研究高性能的新型稀疏模型结构，在模型效果、训练效率、推理效率方面领先 Transformer 等传统模型结构。
（2）基础模型分析：研究模型规模、数据组织方法、模型预测策略等对于不同下游任务的效果影响，对模型在推理计算等复杂任务的能力进行溯源，通过思维链构建、课程学习等方式，提升大模型复杂任务的解决能力。
（3）新型模型对齐算法研究：研究模型对齐技术，包括指令微调、奖励模型、偏好学习、强化学习、少样本低参数微调等技术的优化，以及 SuperAlignment 等新型对齐方法，提升对齐模型能力。
（4）多模态预训练：研究多模态统一表征建模的方法，探索模型的跨模态知识迁移能力。通过不同模态融合学习，增强基础模型在语言模态任务、视觉模态任务、音频模态任务、以及跨模态任务的模型效果。
验收标准：基于飞桨研究基础模型训练方法，在 20 个以上典型应用任务上，效果有显著提升，具备实际应用价值，并支持大模型云平台应用。

3.2 智能体

概述：研究面向智能体（Agent）的技术架构、能力评估和优化算法，推动智能体的发展和应用，对未来 AI 技术大规模的应用具有重要价值。
研究方法推荐：
（1）智能体技术架构：研究单智能体、多智能体系统架构，智能体长期记忆机制，智能体之间的协作与通信机制等。
（2）智能体能力评估：提出科学的智能体能力评估方式，覆盖智能体指令遵循、工具规划、反思、长/短期记忆等多种核心能力；建设相关的 benchmark；研究智能体自动评估算法等。
（3）智能体优化算法：研究基于机器学习和强化学习的智能体优化算法，提升智能体的学习效率、效果。重点关注算法的有效性、易用性、可扩展性，以及计算资源利用率和在动态环境中的稳定性。
验收标准：
（1）设计智能体技术架构，在多种任务场景上具备较好的通用性、灵活性，并保障任务效果。
（2）设计全面的智能体能力评估标准，产出高质量 benchmark。智能体自动评估效果达到实际可用标准。
（3）提出智能体优化算法，具备有效性、灵活性、低成本等特点，在多种任务场景上均能产生明显效果。

3.3 数据建设

概述：数据建设在大语言模型的指令微调和对齐训练中发挥着至关重要的作用，直接影响着大语言模型是否能够有效地激发能力，且与人类的价值观保持一致。大语言模型的数据构建存在众多值得研究的问题和方向，包括但不限于跨模态对齐数据自动构建、数据质量控制、模型缺陷发现、模型效果自动评估、用户反馈的挖掘和利用等。

研究方法推荐：
（1）模型效果自动评估：研究如何改善大模型批评能力，自动评估大模型的方法和框架，提升相关方法与人类评估的一致性。
（2）数据自动构建、改进与模型的自我提升：研究如何自动发现模型的缺陷，自动建设和改进相关数据，推动模型效果自我迭代。
（3）用户反馈信息的挖掘和利用：研究如何挖掘和利用用户反馈信息，进一步提升大模型的效果。
验收标准：
（1）基于以上研究内容，形成一套能够自动/半自动的执行的方法，并可以有效解决对应的问题。
（2）基于以上研究内容，实现对数据的优化，对应的评估集合上效果有显著提升（WinRate 等指标提升 10%以上）。

4. 计算机视觉

本课题希望对深度学习技术在计算视觉中的前沿问题和产业化实践中的实际难题进行分析和探索。在前沿问题探索中，着重方法的创新性、领先性和可推广性，鼓励探索基础模型的突破，填补视觉技术领域的空白。在实际产业化难题解答中，针对实际场景，着重方法的效果和效率，促进产业化的高效、高质量发展。
建议研究方向：

4.1 文档图像版面分析关键技术

概述：版面分析技术，作为文档处理与理解的重要组成部分，专注于对文档版面结构进行自动定位和解析。它旨在将复杂的文档图像分解为逻辑上独立且易于处理的元素，如文本、图像、表格、标题、页眉页脚等，并确定这些元素之间的空间关系和层次结构。随着数字化文档的快速增长，版面分析技术对于提高文档检索效率、实现自动化内容提取和文档重排等任务至关重要。期望能够深入研究版面元素的精确定位和解析方法，探索多源信息以增强版面分析的准确性，以应对复杂多变的文档版面；针对不断增长的文档数据和动态变化的版面结构，研发高效高精度的训练策略。结合多模态方法，寻求融合方案，进一步提升版面分析效果。
研究方法推荐：
（1）基于图模型、层次聚类等算法的版面结构解析方法，融合文档中的逻辑结构和层次关系增强定位准确性，实现准确的版面解析。
（2）基于增量学习，允许模型在不断接收新的文档数据和动态变化的版面结情况下持续学习，从而适应数据分布的变化和新的任务需求，实现高效高精度的训练策略。
（3）基于多模态的版面分析方法，进一步提升效果。
验收标准：
（1）基于飞桨，完成版面分析等方向的 SOTA 算法，显著提升现有模型精度。

（2）基于飞桨，打造高效的模型训练系统，能够在接收新的文档数据和动态变化的版面结情况下高效高精度学习。
（3）基于飞桨，基于图模型、多模态模型等其他方案的有效融合方法，提升版面分析精度。

4.2 文档图像信息抽取关键技术

概述：文档图像信息抽取技术在多模态任务中扮演着重要的角色，其主要目标是从复杂的图像文档中准确有效地提取所需的键值对。期待深入研究视觉与语义信息的深度融合技术，以实现对文档图像中文本、符号、图形等多模态信息的精准、高效提取以及识别。项目鼓励在多角度的探索中，研发出表格识别和公式识别的精准高效算法；同时可借助多模态或 RAG 等方式，寻找与大模型更好的结合方案。针对训练数据难以获取和标注的现状，本研究方向寻求探索出有效的文本类数据生成算法。这有助于提高从图像文档中获得有价值信息的效率和准确性，从而推动文档数字化以及自动化处理技术的进步。
研究方法推荐：
（1）基于序列识别或基于大模型的表格识别策略，提升表格内容恢复的准确率。
（2）基于 Sequence2Sequence 或 多模态方案的通用公式识别策略，提升公式内容恢复的准确率。
（3）基于 RAG 的内容检索策略，提升信息抽取的准确率和效率。
（4）基于文生图的高效数据生成方案。
验收标准：
（1）基于飞桨，完成文本识别、文本检测、端到端、表格识别、公式识别、多模态大模型等方向的 SOTA 算法，显著提升现有模型精度。
（2）基于飞桨，打造更高效的 RAG 系统，显著多页长文的信息抽取准确率。
（3）基于飞桨，打造文字类的数据生成工具，基于生成数据显著提升文本、表格、公式等场景上的识别效果。

4.3 图像分类模型关键技术

概述：图像分类技术在计算机视觉领域具有广泛应用，其核心目标是将输入的图像准确地归类到预定义的类别中。期望探索并优化深度学习模型，以提高图像分类的准确性和效率。鼓励从网络架构设计、模型优化、数据增强、多模型融合等多个角度入手，全面提升图像分类算法的性能。同时，针对训练数据不足或类别不均衡的问题，本研究方向也寻求有效的数据扩充和类别平衡策略。
研究方法推荐：
（1）研究并设计高效的 CNN 架构、Transformer 架构、Mamba 架构等，以提升图像特征的提取能力。
（2）探索模型蒸馏、剪枝等模型压缩技术，以在保持性能的同时降低模型复杂度。
（3）借助大模型零样本识别能力，针对数据不均衡问题，采用过采样、欠采样或 StableDifusion 等技术进行类别平衡。
验收标准：

（1）基于飞桨，开发出高性能的图像分类模型，显著提升现有分类准确率。
（2）基于飞桨，实现模型的优化和压缩，确保在不损失性能的前提下，降低模型复杂度和计算资源消耗。
（3）基于飞桨，打造一套高效的数据增强和类别平衡工具，显著提高在数据稀缺或类别不均衡场景下的模型性能。

4.4 目标检测模型关键技术

概述：目标检测技术是计算机视觉领域中的一项重要技术，旨在识别和定位图像中的特定目标对象。本研究方向致力于提高目标检测的精度和速度，从而实现对图像中多个对象的准确识别和定位。为了实现这一目标，鼓励研究人员探索和创新深度学习模型、优化算法，多模型融合等策略以提升目标检测系统的整体性能。
研究方法推荐：
（1）研究和改进现有的目标检测算法，如 PP-YOLO-E、YOLOv10、RT-DETR 等，以提高检测的准确率和速度。
（2）探索多模型融合技术，如目标检测+图像分类的模型融合技术，提升目标检测最终的精度。
（3）借助大模型零样本识别能力，针对数据不均衡问题，进行数据增强，来提升模型效果。
验收标准：
（1）基于飞桨，开发出高性能的目标检测模型，显著提升对图像中目标的检测精度和速度。
（2）基于飞桨，开发出目标检测的多模型融合方案，显著提升对图像中目标的检测精度。
（3）基于飞桨，开发一套高效的数据增强工具，显著提高在类别不均衡场景下的模型性能。

5. AI+科学计算

AI 科学计算是指使用人工智能方法、利用计算机再现、预测和发现客观世界运动规律和演化特征的全过程。通过 AI 学习自然规律、求解数学模型并应用于工程实践和科学探索，解决航空航天、船舶制造、生物计算、地球科学、能源勘探等领域的难题。科学计算中的物理规律，通常使用微分方程或其它数学模型来描述，求解一般面临维数高、计算时间长、计算量大、并行效率低等难题，通过神经网络模拟函数的方法，通常可以简化方程的求解。当前相关的工作包括但不限于物理信息约束神经网络 PINN 方法，数据驱动的傅里叶神经网络操作 FNO 方法，AI 算法与传统数值方法结合算法。本课题期望探索 AI 技术和科学计算任务相结合的创新方法，具备良好的实用性和可推广性，更高效解决各领域的科学计算问题。建议研究方向：研究科学计算相关各领域的 AI 技术解决方案，超越传统方法，包括不限于：

5.1 AI+流体力学：非定常水动力智能预报方法研究

概述：本研究方向旨在突破下一代大型运载工具在超常规恶劣发射条件下安全出水的关键技术。

早期的理论研究往往基于势流理论，这些理论均无法考虑粘性的影响，无法准确预测自由表面和物体附近的流动细节。现有大多研究建立的模型，需要实时求解微分方程，尚无考虑环境变化和运动的空泡流体动力显式模型。随着计算技术的发展，实时耦合求解流场和回转体运动，可详细地获得流场演化对运动特性的影响。然而数值模拟的效率较低，只适用于特定工况或机理研究，无法直接应用于空泡流动实时闭环反馈控制中。近年来，随着人工智能技术的发展，深度学习模型由于训练效率高、预测能力强等特点，不仅成功应用于运载工具运动轨迹在线生成，而且可以推广运用到流体力学的建模或预测中。

针对现有技术在水下发射过程中空泡流动稳定性控制、非定常非线性水动力预报等方面的不足，建议围绕通气空泡稳定控制与非定常水动力预报这一关键科学问题深入开展基础科学研究与技术创新，揭示通气空泡流体动力演化机制，构建非定常水动力表征方法快速预报方法。期望通过本方向研究，推动我国下一代大型运载工具水下发射技术的发展。

研究方法推荐：

（1）通过数值模拟对空泡流动进行预测分析，结合空化水洞实验对模拟结果进行验证和优化。

（2）利用 LSTM 或卡尔曼滤波实现时序压力、水动力的快速预测。

（3）结合实验和计算数据，对独立膨胀原理进行修正，完善和修正通气空泡的回射力模型。

（4）结合理论模型，融合实验和计算数据，采用深度神经网络等机器学习方法，建立物理与数据混合驱动的水动力智能预测模型。

验收标准：

（1）围绕该科学问题，发表高水平 SCI 论文至少一篇。

（2）基于飞桨实现所提出的算法，并贡献到百度组织对应的代码仓库中。

（3）整理并开源论文所使用的数据集。

5.2 AI+气象海洋：基于深度学习的极端天气事件预测和分析

概述：极端天气事件（如台风、暴雨、热浪等）对人类社会和自然环境造成重大影响。传统的预测方法受限于复杂的气候系统和数据的高维度性，难以准确预测这些事件。深度学习技术可以处理大规模数据和复杂非线性关系，有望提升极端天气事件的预测精度。研究目标为开发一种基于深度学习的极端天气事件预测模型，通过多源数据融合，提高极端天气事件的预测精度和提前量。

研究方法推荐：

（1）研究多源数据融合模式，如结合气象观测数据和数值天气预报模式（包含大气物理机制）数据，构建多源数据融合的算法，从而提升网络模型对高维、多通道数据的处理能力。

（2）利用 CNN、GNN、FNO 以及 Attention 等技术，研究能够更好表征高维、多源数据内部特征关联的网络架构。

（3）研究结合大气动力模型和人工智能模型的深度融合范式，将大气动力模型对应的物理机理内嵌到时空网络中，以提高对极端天气和气候预测的准确性。

（4）结合 MoE 中多模型组合的机制，研究气象领域**多个 SOTA 模型**的集成预报。通过集成多个模型的预测结果，更好地捕捉大气系统的复杂性和多变性，从而提高气象预报的准确性和可靠性。

验收标准：

（1）围绕该科学问题，基于飞桨框架，可发表高水平 SCI 论文至少一篇。

（2）以上算法基于飞桨框架实现，并可贡献到百度组织对应的代码仓库中。

（3）整理并开源论文所使用的数据集。

5.3 AI+材料化学：基于深度学习方法的材料属性预测

概述：

新材料的物理和化学特性复杂多变，准确预测其属性，特别是实际合成和使用条件下的属性，是物质科学领域中长期存在的挑战，也是材料工业数字化转型的核心挑战之一。能够在广泛的元素、温度和压力范围内实现准确高效的材料模拟与性质预测，为材料设计的数字化转型提供了强有力的支持。新材料探索对纳米电子学、能量储存和医疗健康等多个领域的技术进步至关重要。材料设计中的一个核心难点是如何在不进行实际合成和测试的情况下预测材料属性。

研究方法推荐：

（1）融合主动学习、分子动力学模拟或密度泛函等技术，构建高效的数据生成方案。

（2）基于图神经网络、卷积网络等深度学习技术预测材料在不同压力或温度下在原子层面的能量、力和应力或者可用于构造分子动力学的势函数模型，衔接起传统分子动力学（LAMMPS 等）模拟软件。

验收标准：

（1）围绕该科学问题，发表高水平 SCI 论文至少一篇。

（2）基于飞桨实现所提出的算法，并贡献到百度组织对应的代码仓库中。

（3）整理并开源论文所使用的数据集。

5.4 生物计算：基于语言模型注意力的蛋白质词表技术用于蛋白设计

概述：尝试建立基于蛋白词表的蛋白设计新方法，打开蛋白语言模型的黑箱，构建通用和专属的蛋白功能团词表，设计全新功能的小蛋白、无结构蛋白、预测蛋白未知功能等，以更好的指导蛋白质设计，填补目前蛋白设计方法的不足。

研究方法推荐：

（1）蛋白语言模型的注意力头

尽管大语言模型（简称大模型）在各个领域取得了前所未有的突破，人类对大模型工作机制的理解仍旧处于初级阶段。但蛋白质则完全不同，人类仅对极其有限的蛋白功能团做过研究。类比于小分子中的“官能团”概念（如羟基、羧基等），假设蛋白质中也存在特定序列组成的相对稳定的序列结构（但不一定会折叠成稳定的 3D 结构）。开发一种注意力头机制，用于生成蛋白质中的特定序列。

（2）词表提取

设计一种词表提取的算法，可以将注意力头中的信息转化，从而能够解决不连续词的提取问题，不连续词能辅助理解蛋白的远端调控机制。需要指出，通过直接切割序列的方式提取词表无法获得不连续词。

（3）评估数据集构建

需要构建一套用于评估算法的数据集，此前尚无本领域可参考的数据。用于评估词表提取算法的覆盖率、准确率。

数据集需要包含：1）不少于 30 个残基功能标注的蛋白序列（用于评估词的覆盖率）；2）功能肽（用于评估连续词准确率）；3）抗体；4）蛋白功能（GO terms）；5）蛋白结合位点；6）翻译后修饰等。

（4）算法可以应用于以下应用场景：

1)设计全新功能小蛋白；

2)设计无结构蛋白（IDP）,理解蛋白质相变（LLPS）；

3)预测未知功能结构域（DUF）；

4)理解蛋白的远端调控机制（也称变构调控）。

验收标准：

（1）包含不少于 10 个词表评估数据集。

（2）蛋白词表算法，能够处理 Uniref50 所有数据。

（3）构建蛋白质词表，词表对于评估数据集的平均覆盖率达到 80%。

（4）发表高水平 SCI 论文一篇。

6.大模型推理部署关键技术研究与开源实现

本课题希望以生成式大模型为切入点，结合模型结构设计、量化压缩、推理优化、服务部署等端到端设计和优化，深如研究大模型推理部署过程中的如超大规模多专家模型、超长上下文推理、多精调模型集约化部署、并发投机采样优化等推理部署关键技术，极致降低推理部署过程中的成本。研究过程中应注重创新性、实用性和可拓展性，突破对应场景的性能瓶颈，同时促进开源生态高效快速发展。

建议研究方向：

6.1 万亿参数级多专家大模型高性能推理部署优化

概述：混合专家(MoE, Mixture of Experts)模型是当前大语言模型重点关注的方向之一，通过不同领域的专家组合，即能结合各领域的知识得到更好的模型效果，同时也能方便的通过增减专家数量实现可扩展性。然而随着专家个数的增加，模型参数量会不断变大，万亿参数级别的多专家模型也成为大模型发展的趋势之一。本方向期望研究 MoE 模型精度无损极致量化、多机高效并行推理、多专家异构并行推理等技术，实现万亿参数级别多专家模型的高效稳定部署及极致性能优化。

研究方法推荐：联合压缩、推理设计量化方式和算法、实现精度无损量化和极致推理性能加速兼顾；通过多机并行、专家并行、异构并行等优化方式，充分利用机器资源，实现高效并行推理加速

验收标准：基于飞桨核心框架、PaddleNLP 大模型套件，形成多专家模型的无损量化、多机并行和异构并行的通用方案，实现万亿参数级别的多专家模型高性能推理服务部署通用方案

6.2 超长上下文高性能推理优化

概述：长上下文作为生成式大模型的重点方向之一，但随着上下文长度的增加，有着显存占用增加、计算量指数级增长的难点。期望基于飞桨核心框架、PaddleNLP 大模型套件等，研究超长上下文的通用高性能推理优化方案，结合细粒度算子融合优化、KV Cache 低比特量化，上下文 Chunk 划分，稀疏化推理等方法，实现超长上下文推理支持及性能领先

研究方法推荐：

（1）结合 PaddleNLP 的细粒度算子融合功能，对长上下文下的关键算子，如 Multi-Block Attention 和 Quant/Dequant 等，进行深度融合优化，以提升推理速度。

（2）针对长上下文中 KV Cache 占用显存量大的问题，通过 4bit、2bit 等极致量化优化，节省推理显存开销。

（3）针对长上下文推理具有一定的稀疏性，从 KV Cache、Token、Layer 等多个维度进行稀疏化推理探索。

验收标准：基于飞桨支持 1024K 上下文推理，并进行推理性能极致优化，性能持平或领先业界主流生成式大模型推理框架。

6.3 大模型集约化高性能推理部署方案

概述：随着大模型发展，逐渐赋能到不同的领域、业务场景，需要根据不同场景的知识、信息、数据做精调，LoRA 成为低成本精调的主流方案之一，而在大量业务的不规律请求的背景下，实现 LoRA 的集约化高性能推理部署是一种有效的整合资源、降低部署成本的方式。本研究方向期望基于飞桨核心框架、PaddleNLP 大模型套件等，研究集约化 LoRA 部署的高性能解决方案，结合细粒度算子融合优化、分段 LoRA GEMM 计算、多流异步推理、量化推理等方式，实现集约化 LoRA 推理在性能、显存占用等方面均实现最优。

研究方法推荐：

（1）结合 LoRA 的计算特性和 Encoder/Decoder 的计算瓶颈，实现高性能细粒度融合算子。

（2）通过多 Stream 并发方式掩盖 LoRA 的计算和调度开销，结合压缩推理设计并实现精度无损、性能更优的 LoRA 量化推理方案等。

验收标准：基于飞桨在主流开源大语言模型上支持高性能 LoRA 节约化量化推理部署，推理性能领先业界主流大模型推理框架。

6.4 大语言模型的投机采样推理加速方法研究

概述：大语言模型推理性能慢的因素之一是其需要自回归解码，投机采样推理加速方法，通过引入草稿模型辅助解码，实现大模型的并行解码，从而在不损失生成效果的前 提下显著提高推理速度。本方向旨在研究投机采样的原理、实现技术，并在大语言模型推理加速中取得理想的应用效果。

研究方法推荐：

（1）草稿模型的选择如何使得预测足够轻量且准确，高效的验证策略可以兼顾效率提升和效果无损。

（2）一些低成本训练技术可以进一步优化草稿模型或者目标模型的效果，实现更高的接受率，用以提升加速性能。

验收标准：

（1）创新的投机解码技术，发表高水平论文 1 篇。

（2）基于飞桨实现投机解码高性能推理系统并开源。

7.跨模态大模型技术及 AIGC 应用

概述：跨模态是指在不同类型的数据和模态之间进行交互和融合的技术。跨模态大模型技术创新带来了 AIGC 应用突破，让 AI 在预训练过程中同时学习模态间和模态内的多种关联性，提升“图像”，″视频″ 和“文本”跨模态语义匹配效果；大模型中的核心结构 Transformer 在扩散模型中也逐步展现出了更好的效果和更强的可扩展性，同时驱动应用创新包括文生图、图生图、文生视频等。

建议研究方向：

7.1 大语言模型与其他模态融合

概述：目前跨模态大模型重点将图像模态和大语言模型结合，例如 LLaVA，miniGPT4 等，通过端到端训练的方案有效将视觉特征和语言特征进行对齐，在图像理解、看图创作、图表理解等任务的能力不断提升。随着 MoE（混合专家模型）的兴起，能够进一步提升多模态大模型的性能和计算效率，此外大语言模型和其他模态（例如视频、音频）的融合，实现任意输入输出的多模态大模型，也是业界前沿的研究方向，能够使用更多样的训练数据，提升多模态大模型的能力边界。

研究方法推荐：

（1）MoE 结构在多模态大模型的应用，基础模型可参考 CogVLM。

（2）支持任意模态输入输出的多模态大模型，基础模型可参考 NextGPT。

验收标准：

（1）基于飞桨，在多模态大模型中适配 MoE 结构，优化训练方案，在下游任务中达到 SOTA 效果。

（2）基于飞桨，训练支持多种模态的多模态大模型结构，输入支持三种模态以上，输出支持两种模态以上。

7.2 基于 Transformer 架构的扩散模型

概述：扩散模型是 AIGC 领域核心的模型架构，在文生图方向已经大量投入使用，业界扩散模型以基于卷积 UNet 为主。随着 Stable Diffusion 3 的出现，基于 Transformer 架构的扩散模型能力被大幅挖掘，具备更强的可扩展性，参数规模也可以达到 10 亿以上。针对 Transformer 架构的扩散模型，值得更深入的分析 scaling law 以及更优的文本图像融合方案。另外基于卷积的扩散模型可以实现多样的小样本定制化训练方案，并提升文生图的可控性，对于 Transformer 架构的扩散模型同样需要这类训练方案的探索。

研究方法推荐：

（1）基于 Transformer 架构的扩散模型 scaling law 分析，分析扩散模型下数据规模、参数规模、训练资源和模型性能的关系。

（2）基于 Transformer 架构的扩散模型，实现可控生成和小样本训练方案。

验收标准：基于飞桨，设计更优的扩散模型结构，结合飞桨分布式训练能力，验证提出的训练方案在公开评测集上达到 SOTA 效果。

7.3 文生视频应用

概述：在文生图扩散模型取得技术突破和巨大进步后，文生视频是 AIGC 下一个亟待突破的应用方向。随着 SoRA 的出圈，急需结合当前文生图模型的技术方法，扩展应用到视频生成领域。针对低资源和大规模不同场景，实现视频生成质量的显著提升，具有很强的研究价值与实际意义。

研究方法推荐：文生视频底座模型优化，基础模型可参考 DiT。

验收标准：基于飞桨，提出更高效的文生视频底座模型，生成时长超过 10s，支持任意分辨率的视频。

7.4 扩散模型推理效率提升方法研究

概述：扩散模型结构在文生图、文生视频等领域展现出卓越的效果。然而，在推理阶段面临巨大的计算开销，高分辨率图像、长时间序列视频场景下计算量急剧上升，这在一定程度上限制了其在实际应用中的广泛部署。因此，研究如何提升扩散模型的推理效率具有重要的理论意义和实践价值。本方向旨在探索扩散模型推理效率提升的方法，通过优化模型结构、采样方法、模型压缩等手段，降低扩散模型的计算复杂度，提高其在各种应用场景下的推理速度和资源利用率。

研究方法推荐：高效模型结构、采样算法、跨步蒸馏、模型量化压缩、分布式并行推理等。

验收标准：

（1） 创新的扩散模型加速方法，发表高水平论文 1 篇。

（2） 基于飞桨实现新提出的优化方法并开源。

8. 基于大模型重点领域数据与应用

概述：本议题致力于促进数据共享与合作，形成丰富的行业数据生态，面向重点行业领域（如金融、法律、医疗）进行专业数据集建设。在此基础上，基于行业

核心场景开发高价值工具与应用，设立明确的评测标准并对所提供的数据集进行验证，从而验证数据集的质量及其对模型能力提升的有效性。

研究方法推荐：

（1）构建专业领域数据集；

（2）创建行业领域应用；

（3）设立评测标准并进行评测。

验收标准：

（1）数据集构建与合作成果：与不少于 1 个机构或企业合作，构建 1 套专业领域高质量数据集，数据形态不少于 3 个模态（包括文本、图片、声音、视频等），规模不少于 5 万条。

（2）应用功能实现：基于文心系列大模型，进行领域数据大模型行业工具或 Agent 设计与开发。开发的工具或 Agent 应用需具备至少 3 项针对行业领域的核心功能，且功能实现符合预期，能够在实际场景中稳定运行。

（3）评测结果达标：使用新构建的数据集对基座模型进行训练后，依据制定的评测标准进行全面评测，能够使模型在关键性能指标（如准确率、召回率等）上相较于引入数据前有明显提升，从而证明数据集的有效性及模型能力的提升。