大模型的软硬件协同优化和高效部署技术

清华大学

一、转化对象

上海无问芯穹智能科技有限公司

二、服务机构

清华大学科研院、华控技术转移有限公司

三、转化特色

普通许可+增资入股，先许可后入股

四、案例简介

清华大学电子系汪玉教授团队研究并提出“面向大模型的软硬件协同算力优化技术”，利用模型、算法、系统与硬件的跨层协同优化，实现面向异构算力的全栈式优化，可使模型训练和推理速度提升 1 个数量级以上，在提升性能、降低成本、缩短开发周期等方面具有显著优势。已申请相关发明专利 40 余项，其中已授权专利 20 余项。

为适应 AI 领域快速发展特点，该项目采取“先许可后入股”模式进行转化，学校先后许可并增资入股上海无问芯穹智能科技有限公司。在学校及科研团队的支持下，公司发展迅速，截至 2025 年 3 月已完成多轮融资，融资额近 10 亿元，并推出AI 基础设施产品 Infini-AI 异构云平台等，应用于多家大模型公司客户。

五、转化过程

该项目采取“先许可后入股”模式进行转化，由具有丰富研发和市场经验的清华校友等作为主要创业团队，注册成立无问芯穹公司，学校先以技术许可方式授权企业使用，待项目发展到一定阶段后，再将相关成果通过作价的方式增资入股，较好地平衡了领域特点及团队需求等，大大加快了成果转化进程。2023 年 5 月，无问芯穹公司注册成立；12 月，学校将部分专利许可公司使用；2025 年 3 月，学校进一步将该成果作价 增资入股至公司。公司获得相关技术后，已成功应用于主营产品中的无穹 AI 云平台和端侧大模型推理专用 LPU IP，不仅提升了产品性能与资源利用率，还加速推动了公司与多家客户的商业合作落地，为公司创造了显著且可持续的商业价值。

六、转化效益

该项目取得了良好的转化收益。项目公司发展迅速，截至 2025 年 3月已完成多轮融资，融资额近 10 亿元。依托“多元异构、软硬协同”的核心技术优势，无问芯穹打造了连接“M 种模型”和“N 种芯片”的“M × N”AI 基础设施新范式，实现多种大模型算法在多元芯片上的高效协同部署，助力实现普惠AI，为千行百业注入新质生产力。

七、成果完成人及团队

汪玉教授，清华大学电子系主任，IEEE Fellow，国家自然科学基金杰出青年基金获得者，清华大学信息科学技术学院副院长、天津电子信息研究院院长。担任中国电子学会第十一届理事会常务理事，中国电子学会青年工作委员会主任委员，全国电子信息学科建设（推进）委员会主任。