English version

This is a **detailed research plan** for the year 2025, focusing on the two new core areas: *small language models (SLMs)* and *algorithm performance optimization*. This plan is tailored to meet the ambitious KPI of achieving **5 patents** or **10 SCI papers**.

2025 Research Plan for AI R&D

1. Objective

The research will focus on:

- **Small Language Models (SLMs):** Exploring lightweight, efficient, and deployable language models for edge devices or embedded AI.
- **Algorithm Performance Optimization:** Developing methods to reduce computational complexity and improve processing speed on limited hardware.

The key deliverable is to achieve **5 patents** or **10 SCI-indexed publications** in these areas.

---

2. Research Breakdown by Quarters

Q1 2025: Laying the Foundation

Focus: Literature review, resource setup, and identifying key research problems.

Task	Details	Deliverables
2.1 Comprehensive Survey	Conduct a literature review on current SLMs (e.g., DistilBERT, TinyBERT) and optimization methods (e.g., quantization, pruning, low-rank factorization). Identify gaps in current technologies.	Technical survey report and gap analysis.
2.2 Resource and Tool Setup	Set up a research environment with tools like PyTorch, TensorFlow, ONNX for lightweight model deployment, and libraries for optimization (e.g., TensorRT, OpenVINO).	Functional development environment.
2.3 Problem Statement Formulation	Narrow down specific challenges (e.g., efficient SLM architecture design, latency improvements on RTX 3090). Define 2-3 key research problems.	Finalized research problems for Q2.

Milestone:

- Complete literature review.
- Finalize 2-3 focused research questions for each topic.

Q2 2025: Model Development and Experimentation

Focus: Prototype design, experimentation, and initial model training.

Task	Details	Deliverables
2.4 Develop Small Language Models	Design lightweight Transformer architectures with optimizations like quantization (8-bit/4-bit), knowledge distillation, and pruning. Experiment with datasets (e.g., WikiText, Common Crawl).	Prototype SLM with benchmark results.
2.5 Optimize Existing Algorithms	Apply and compare optimization techniques: - Matrix decomposition. - Kernel fusion. - Low-rank approximations. Test efficiency on RTX 3090 GPU.	Optimized algorithms and benchmarks.
2.6 Benchmark Models	Evaluate accuracy, latency, and memory consumption of SLM prototypes and optimized algorithms using established metrics (e.g., FLOPs, inference time).	Comparative performance report.

Milestone:

- Prototypes of small language models and optimized algorithms.
- Preliminary results ready for internal review and feedback.

Q3 2025: Patent Filing and Publication Preparation

Focus: Consolidate results, file patents, and write scientific papers.

Task	Details	Deliverables
2.7 Patent Drafting and Submission	File patents for innovative SLM architecture and algorithm optimization methods. Focus on uniqueness and feasibility.	2-3 patent applications submitted.
2.8 Writing and Submitting SCI Papers	Prepare 4-5 research papers for SCI-indexed journals, such as: - *Journal of Artificial Intelligence Research. - IEEE Transactions on Neural Networks. - Pattern Recognition Letters.*	4-5 SCI submissions.
2.9 External Collaboration	Collaborate with academic partners or external researchers for validation or feedback on results.	Joint research validation.

Milestone:

- Submit 2-3 patent applications.
- Submit 4-5 SCI-indexed papers for review.

Q4 2025: Final Optimization and Reporting

Focus: Refine solutions, finalize patents/papers, and prepare technical presentations.

Task	Details	Deliverables
2.10 Final Model Optimization	Optimize developed SLMs further for real-world deployment. Perform extensive testing for scalability and robustness.	Final, deployable lightweight models.
2.11 Patent Resubmission and Final Edits	Address reviewer feedback for pending patents and journal submissions. Finalize all outstanding applications.	Additional 2-3 patents finalized.
2.12 Research Showcase	Present findings in company-wide meetings, focusing on advancements and potential applications for production.	Technical presentations and internal reports.

Milestone:

- File 5 patents (total).
- Publish 8-10 SCI-indexed papers.
- Present final results to the leader's company.

---

3. Timeline Overview

Quarter	Key Milestones
Q1	Finalize research questions and setup.
Q2	Develop and benchmark prototypes.
Q3	File 2-3 patents and submit 4-5 SCI papers.
Q4	Finalize patents/papers, optimize models, and present results.

---

4. Resource Requirements

To achieve the above goals, the following resources are essential:

1. **Computational Hardware:** Access to at least 2 RTX 3090 GPUs (or equivalent) for training and inference.
2. **Tools and Libraries:** PyTorch, TensorRT, OpenVINO, ONNX, and access to benchmark datasets.
3. **Collaboration:** Support for academic collaboration or external validation.

---

5. Expected Outcomes

- **5 Patents Filed** (covering innovative architectures, optimization techniques).
- **10 SCI Papers** published in high-impact journals.
- **Deployable SLMs:** Lightweight models ready for use on edge devices.
- **Performance Optimizations:** Proven methods to reduce latency and computational cost on GPUs.

Vietnamese version

Đây là **kế hoạch nghiên cứu chi tiết** cho năm 2025, tập trung vào hai lĩnh vực mới và cốt lõi: **mô hình ngôn ngữ nhỏ** và **tối ưu hóa hiệu suất thuật toán**. Kế hoạch này được xây dựng để đáp ứng KPI đạt **5 bằng sáng chế** hoặc **10 bài báo SCI**.

Kế Hoạch Nghiên Cứu AI R&D Năm 2025

1. Mục Tiêu

Tập trung nghiên cứu vào:

- **Mô hình ngôn ngữ nhỏ:** Khám phá các mô hình ngôn ngữ nhẹ, hiệu quả và có thể triển khai trên thiết bị biên hoặc AI nhúng.
- **Tối ưu hóa hiệu suất thuật toán:** Phát triển các phương pháp giảm độ phức tạp tính toán và tăng tốc xử lý trên phần cứng hạn chế.

Mục tiêu chính là đạt được **5 bằng sáng chế** hoặc **10 bài báo được SCI-index**.

---

2. Chi Tiết Kế Hoạch Theo Quý

Quý 1/2025: Đặt Nền Móng

Trọng tâm: Tổng quan tài liệu, chuẩn bị tài nguyên và xác định vấn đề nghiên cứu.

Nhiệm vụ	Chi tiết công việc	Kết quả đầu ra
2.1 Khảo sát tổng quan	Nghiên cứu các mô hình ngôn ngữ nhỏ (ví dụ: DistilBERT, TinyBERT) và các phương pháp tối ưu hóa (như lượng tử hóa, rút gọn mô hình). Xác định khoảng trống công nghệ.	Báo cáo tổng quan kỹ thuật và phân tích khoảng trống.
2.2 Cài đặt môi trường	Xây dựng môi trường nghiên cứu với các công cụ: PyTorch, TensorFlow, ONNX, và thư viện tối ưu như TensorRT, OpenVINO.	Môi trường phát triển hoàn chỉnh.
2.3 Xác định bài toán	Xác định rõ 2-3 vấn đề cốt lõi cần nghiên cứu (ví dụ: thiết kế kiến trúc mô hình SLM, cải thiện độ trễ trên GPU RTX 3090).	Danh sách vấn đề nghiên cứu.

Cột mốc:

- Hoàn thành tổng quan tài liệu.
- Xác định 2-3 câu hỏi nghiên cứu trọng tâm.

Quý 2/2025: Phát Triển và Thử Nghiệm

Trọng tâm: Xây dựng mô hình thử nghiệm và đánh giá hiệu suất ban đầu.

Nhiệm vụ	Chi tiết công việc	Kết quả đầu ra
2.4 Phát triển SLM	Thiết kế kiến trúc Transformer nhẹ với các kỹ thuật tối ưu như lượng tử hóa (8-bit/4-bit), distillation, và pruning. Huấn luyện trên các bộ dữ liệu như WikiText, Common Crawl.	Nguyên mẫu mô hình SLM và kết quả đánh giá.
2.5 Tối ưu hóa thuật toán	Áp dụng và so sánh các kỹ thuật tối ưu như: - Phân rã ma trận. - Kernel fusion. - Xấp xỉ hạng thấp. Kiểm thử trên GPU RTX 3090.	Thuật toán đã tối ưu hóa với kết quả benchmark.
2.6 Đánh giá mô hình	Đo lường độ chính xác, độ trễ và hiệu suất bộ nhớ của SLM và thuật toán tối ưu. Các chỉ số như FLOPs, thời gian suy luận sẽ được đánh giá.	Báo cáo so sánh hiệu năng.

Cột mốc:

- Phát triển nguyên mẫu mô hình và thuật toán tối ưu.
- Có kết quả thử nghiệm sơ bộ để đánh giá.

Quý 3/2025: Nộp Bằng Sáng Chế và Xuất Bản

Trọng tâm: Tập trung viết báo cáo, nộp bằng sáng chế và bài báo khoa học.

Nhiệm vụ	Chi tiết công việc	Kết quả đầu ra
2.7 Soạn thảo và nộp bằng sáng chế	Đăng ký 2-3 bằng sáng chế liên quan đến kiến trúc SLM và phương pháp tối ưu hóa mới. Tập trung vào tính độc đáo và khả thi.	2-3 hồ sơ bằng sáng chế được nộp.
2.8 Xuất bản bài báo SCI	Chuẩn bị và gửi 4-5 bài báo khoa học tới các tạp chí SCI như: - *Journal of Artificial Intelligence Research. - IEEE Transactions on Neural Networks.*	4-5 bài báo được gửi đi.
2.9 Hợp tác và phản biện	Hợp tác với các đối tác học thuật để kiểm chứng kết quả hoặc nhận phản hồi cải tiến.	Nghiên cứu chung với đối tác.

Cột mốc:

- Nộp 2-3 bằng sáng chế.
- Gửi 4-5 bài báo SCI-index.

Quý 4/2025: Hoàn Thiện và Báo Cáo

Trọng tâm: Tối ưu mô hình cuối cùng, hoàn thiện tài liệu và trình bày kết quả.

Nhiệm vụ	Chi tiết công việc	Kết quả đầu ra
2.10 Tối ưu hóa cuối cùng	Hoàn thiện và tối ưu mô hình SLM để triển khai thực tế. Thử nghiệm tính ổn định và mở rộng.	Mô hình ngôn ngữ nhỏ sẵn sàng triển khai.
2.11 Hoàn thiện bằng sáng chế và bài báo	Chỉnh sửa và phản hồi ý kiến từ hội đồng đánh giá bằng sáng chế và tạp chí. Nộp hồ sơ bổ sung nếu cần.	2-3 bằng sáng chế và 8-10 bài báo hoàn tất.
2.12 Báo cáo và trình bày	Tổng hợp kết quả và trình bày tại các cuộc họp nội bộ công ty. Tập trung vào tính ứng dụng của nghiên cứu.	Báo cáo kỹ thuật và trình bày kết quả.

Cột mốc:

- Nộp đủ **5 bằng sáng chế**.
- Xuất bản **8-10 bài báo SCI**.
- Trình bày kết quả nghiên cứu với lãnh đạo công ty.

---

3. Tổng Quan Kế Hoạch Thời Gian

Quý	Cột mốc quan trọng
Q1	Xác định câu hỏi nghiên cứu và xây dựng môi trường.
Q2	Phát triển nguyên mẫu và thử nghiệm.
Q3	Nộp 2-3 bằng sáng chế và gửi 4-5 bài báo SCI.
Q4	Hoàn thiện 5 bằng sáng chế và 8-10 bài báo.

---

4. Yêu Cầu Tài Nguyên

Để thực hiện kế hoạch này cần các tài nguyên sau:

1. **Phần cứng tính toán:** Ít nhất 2 GPU RTX 3090 (hoặc tương đương).
2. **Công cụ và thư viện:** PyTorch, TensorRT, OpenVINO, ONNX, và bộ dữ liệu benchmark.
3. **Hợp tác:** Hỗ trợ từ các đối tác học thuật và phản biện bên ngoài.

---

5. Kết Quả Dự Kiến

- **5 Bằng Sáng Chế** được nộp và phê duyệt.
- **10 Bài báo khoa học SCI** được xuất bản.
- **Mô hình ngôn ngữ nhỏ**: Mô hình AI nhẹ, hiệu quả sẵn sàng triển khai.
- **Tối ưu hóa thuật toán:** Các phương pháp tăng tốc và giảm chi phí tính toán.
