Introduction: SemiKong

# The First Industry-Specific Large Language Model for Semiconductor Manufacturing

- **What is SemiKong?**
  - First specialized LLM tailored for the semiconductor industry
  - Foundation model with deep understanding of semiconductor processes
  - Available in 8B and 70B parameter versions

- **Why was it needed?**
  - Generic LLMs lack specialized knowledge for semiconductor challenges
  - Complex physics and chemistry of semiconductor devices require expertise
  - Industry-specific terminology and process flows need dedicated training

- **Key Achievement:**
  - Outperforms larger general-purpose LLMs (GPT-3.5, Claude variants)
  - Surpasses commercial products in expert-level metrics
  - Expert-level understanding of semiconductor manufacturing

- **Industry Impact:**
  - Enables AI-driven solutions for semiconductor manufacturing
  - Foundation for company-specific proprietary models
  - Bridges gap between AI researchers and domain experts

Broader Scope of SemiKong: Motivation

- **Objective:** SemiKong is designed to comprehensively support the **entire semiconductor manufacturing lifecycle**, extending beyond isolated tasks or single process stages.
- Covers both major stages of semiconductor manufacturing:
  - **Front-End-of-Line (FEOL):** fabrication steps performed on the silicon wafer to define and build individual transistor structures, including lithography, etching, doping, and thin-film deposition.
  - **Back-End-of-Line (BEOL):** subsequent steps that create metal interconnects, vias, and packaging to connect the transistors into functional circuits.
- Overcomes the gap between AI researchers (strong in ML but weak in domain knowledge) and semiconductor experts.
- Developed in close collaboration with domain experts to:
  - Systematically structure semiconductor knowledge.
  - Ensure that no critical processes are overlooked.
  - Enable efficient training and evaluation of domain-specialized language models.
- Serves as a foundation for creating:

  - Expert-level LLMs for specific manufacturing stages.
  - Benchmarks for evaluating domain-specific and general-purpose models.

Broader Scope of SemiKong: Ontology Structure

- The ontology systematically organizes the semiconductor manufacturing process into **10 major process groups:**
  1. Substrate Preparation (Wafer Manufacturing, Polishing, Cleaning)
  2. Film Formation (Oxidation, Deposition, Epitaxial Growth)
  3. Patterning (Lithography, Etching)
  4. Doping (Ion Implantation, Diffusion, In-situ Doping)
  5. Planarization (Chemical Mechanical, Etchback)
  6. Cleaning and Surface Preparation (Wet, Dry, Advanced)
  7. Thermal Processing (Annealing, Oxidation, Dopant Activation)
  8. Metrology and Inspection (Physical, Electrical, Defect)
  9. Advanced Modules (High-k/Metal Gate, Strain Engineering, 3D)
  10. Back-End Processes (Interconnect, Metallization, Packaging)
- **Comprehensive coverage spans FEOL and BEOL processes:**
  - FEOL: Active device creation (transistors, gates)
  - BEOL: Interconnection and packaging systems
- Each process group is further **hierarchically decomposed:**
  - Process Group → Process Module → Process Unit
  - Example: *Patterning* → *Etching* → *Reactive Ion Etching* → *Deep Reactive Ion Etching*
  - **Etching alone includes 9 specialized techniques** (Wet, Dry, Plasma, RIE, DRIE, etc.)
- **Expert-developed with industry collaboration:**
  - Created with semiconductor experts from Tokyo Electron Ltd
  - Top-down approach ensuring no critical processes overlooked
  - Serves as benchmark for future general intelligence models
- Enables precise understanding, training, and evaluation of models at any desired level of detail.

# Key Contributions (Detailed, with Full Terminology)

- **1 SemiKong-Corpus: Industry-Specific Knowledge Base**
  - **Scale & Coverage:**
    - 525.6 million tokens, entirely domain-specific.
    - Covers entire semiconductor manufacturing lifecycle:
      - **Front-End-of-Line (FEOL):** wafer preparation, lithography, etching, doping.
      - **Back-End-of-Line (BEOL):** metallization, planarization, interconnects, packaging.
    - Emphasis on etching technology.
  - **Sources:**
    - 129 books and book chapters.
    - 708 etching-specific peer-reviewed papers.

    * 20,000+ research papers spanning manufacturing, defect detection, and design methodologies.
    * 50,000 instruction–response pairs reflecting real-world semiconductor process scenarios.
  - **Processing Pipeline:**
    * Raw PDFs processed via PyPDF to extract text.
    * Cleaned and semantically normalized using GPT-4o-mini, preserving:
      · Tables, equations, hierarchical structure.
    * Converted into structured Markdown for:
      · High-fidelity domain representation.
      · Elimination of noise and misaligned tokens found in generic corpora.

- **2 SemiKong-Trainer: Specialized Foundation Model Pipeline**
  - **Architecture Overview:**
    * Based on Meta's **LLaMA-3 (8B and 70B)** checkpoints.
    * Domain-augmented strategy combining knowledge-rich pretraining with instruction-following fine-tuning.
  - **Core Natural Language Processing Techniques:**
    * **Tokenization:** Byte-Pair Encoding (BPE) via Tiktoken for compact subword representation.
    * **Positional Encoding:** Rotary Position Embedding (RoPE), superior for modeling long-range dependencies.
  - **Fine-Tuning Methodology:**
    * Supervised Fine-Tuning (SFT) directs the model to actionable reasoning, structured dialogue, and domain Q&A.
  - **Post-Training Optimization:**
    * **Low-Rank Adaptation (LoRA):** lightweight domain adaptation while preserving general capabilities.
    * **GPTQ:** post-training quantization for reduced memory footprint and faster inference.

- **3 SemiKong-Eval: Expert-In-The-Loop Evaluation Framework**
  - **Why?**
    * Conventional evaluation metrics (e.g., BLEU, ROUGE) and non-expert annotators fail in expert domains.
  - **Pipeline Components:**
    * Expert-created semiconductor ontology:
      · Covers process groups, modules, and units (e.g., Wet Etching → Plasma Etching → Reactive Ion Etching (RIE)).
    * Benchmark dataset:
      · Easy: 100 questions
      · Medium: 737 questions
      · Hard: 150 questions
    * Evaluation metrics aligned with expert standards:
      · Clarity & Directness (C&D)
      · Practicality & Immediate Usability (PIU)
      · Efficiency & Brevity (E&B)
      · Logical Flow & Coherence (LFC)
      · Expert-to-Expert Communication (EEC)
      · Use of Examples & Specificity (UES)

  - **Iterative Refinement:**
    * Experts annotate outputs with justifications → researchers distill criteria → LLM evaluators align → iterative improvements enhance benchmarks.

- **Data Curation & Pretraining (Detailed)**
  - **Pretraining Dataset:**
    * Designed for depth over breadth.
    * Domain knowledge from:
      · Process manuals, etching-focused physical and chemical research, patents, technical standards.
    * Processed:
      · PyPDF extraction.
      · GPT-4o-mini post-processing for:
      · Fixing structural issues (broken lines, misparsed tables).
      · Normalizing into machine-readable Markdown.
  - **Instruction Dataset:**
    * 50,000 instruction–response pairs:
      · 5,000 explaining semiconductor principles.
      · 5,000 mathematically rigorous etching problems.
      · 40,000 addressing routine control and defect diagnosis.
    * Answers generated via:
      · GPT-4o: conceptual and procedural reasoning.
      · GPT-o1-preview: mathematically complex reasoning.

- **Model Architecture & Training (Detailed)**
  - **Foundation Models:**
    * Meta's LLaMA-3 (8B and 70B), chosen for strong baselines and compatibility with parameter-efficient fine-tuning.
  - **Training Strategy:**
    * Stage 1: Domain-specific pretraining.
    * Stage 2: Supervised Fine-Tuning (SFT) for natural-language alignment to semiconductor scenarios.
  - **Post-Processing:**
    * Low-Rank Adaptation (LoRA): efficient fine-tuning and integration.
    * GPTQ: Post-Training Quantization method for Generative Pre-trained Transformers: quantized weights for lower memory and faster inference.
  - **Hardware & Compute Resources:**
    * SemiKong-8B: 4× NVIDIA A100 80GB, ∼150 hours (15 runs).
    * SemiKong-70B: 8× NVIDIA A100 80GB, ∼200 hours (2 runs).

# 1 Project Overview

**Vision-Language Fine-Tuning with Synthetic Annotations for SEM/TEM Nanomaterial Understanding**

- Leverage publicly available SEM (Scanning Electron Microscopy) and TEM (Transmission Electron Microscopy) image datasets of semiconductor nanomaterials, which typically lack associated text labels or annotations.
- Automatically generate synthetic multimodal annotations — including descriptive captions and Visual Question Answering (VQA) pairs — using powerful pre-trained Vision-Language Models (VLMs).
- Fine-tune state-of-the-art small-to-mid-scale VLMs specifically for nanomaterial domain tasks:
  - Meta LLaMA-3.2 Vision (11B): robust vision-instruction model.
  - Alibaba Qwen-VL / Qwen2.5-VL (3B–7B): open-source, efficient for multimodal tasks.
  - Google PaliGemma-2 (3B & 10B): lightweight, optimized for captioning and VQA.
- The project aims to enable these VLMs to perform expert-level reasoning and understanding of complex nanomaterial imagery.

# 2 Motivation & Goals

- Current SEM/TEM datasets are limited to raw images without textual descriptions or annotations, making them unsuitable for multimodal AI training.
- Manual annotation of nanomaterial data is resource-intensive, subjective, and error-prone.
- Generic VLMs, trained on web-scale general data, are not equipped to handle the unique morphology and terminology of nanomaterials.
- By leveraging synthetic annotations generated automatically by VLMs, we can build a high-quality multimodal dataset and fine-tune models to deliver expert-level analysis.
- The goal is to bridge the gap between general-purpose VLM capabilities and specialized semiconductor nanomaterial knowledge, enabling models to support classification, captioning, and VQA tasks effectively.

# 3 Methodology

**Step 1: Synthetic Annotation Generation**

- Start with publicly available SEM/TEM image datasets of nanomaterials.
- Use a pre-trained VLM to generate:
  - Captions that describe key morphological features, such as grain structure, defects, and textures.
  - VQA pairs that include factual and reasoning-based questions about the image content and answers aligned with domain knowledge.
- The result is a synthetic multimodal dataset pairing each image with rich, domain-relevant text.

**Step 2: VLM Fine-Tuning**

- Fine-tune the selected VLMs using the synthetic multimodal dataset.

- Use instruction-style fine-tuning where each training sample consists of an image and a prompt, with the desired output as the response.
- Employ parameter-efficient adaptation techniques such as LoRA or Q-LoRA to minimize computational overhead.
- Validate performance on a held-out set of real SEM/TEM images.

# 4 Models & Setup

- The project will experiment with three complementary VLMs:
  - Meta LLaMA-3.2 Vision (11B): strong baseline for instruction-following multimodal tasks, compatible with Unsloth/vLLM training pipelines.
  - Qwen-VL / Qwen2.5-VL (3B–7B): open-source and efficient, supporting LoRA/Q-LoRA fine-tuning for multimodal applications.
  - Google PaliGemma-2 (3B & 10B): lightweight, designed specifically for captioning and VQA, with JAX/Flax-based fine-tuning.
- Training involves instruction-tuning on the synthetic dataset using low-rank adaptation, followed by evaluation and refinement.

# 5 Evaluation & Outcomes

- **Evaluation Metrics:**
  - Measure captioning quality with standard metrics: BLEU, ROUGE, and BERTScore.
  - Assess VQA performance through answer accuracy and expert-aligned correctness.
  - Qualitative evaluation by semiconductor domain experts to judge relevance and clarity.
- **Expected Deliverables:**
  - A synthetic multimodal dataset consisting of SEM/TEM images with corresponding captions and VQA pairs.
  - Fine-tuned checkpoints for each of the three VLM families, ready for downstream applications.
  - Evaluation reports, benchmarks, and deployment-ready documentation.
- **Impact:**
  - Enables AI-powered tools for nanomaterial discovery, quality inspection, and scientific documentation.
  - Demonstrates the potential of synthetic multimodal data and VLM fine-tuning in bridging the gap between generic AI models and highly specialized scientific domains.

## 6 How This Project Differs from SemiKong

- **Scope of Data:**
  - SemiKong focuses on textual corpora — curated semiconductor books, papers, patents, and manuals — to train and fine-tune a language-only LLM.
  - This project uses publicly available SEM/TEM image datasets and generates synthetic multimodal annotations (text + image).
- **Modalities:**
  - SemiKong is a pure text-based foundation model, pre-trained and fine-tuned for question answering, explanations, and reasoning in semiconductor process knowledge.
  - This project builds and fine-tunes Vision-Language Models (VLMs) capable of jointly processing images and text for tasks like classification, captioning, and VQA.
- **Synthetic Data:**
  - SemiKong's synthetic data refers to expert-curated instruction–response text pairs within the text domain.
  - This project generates synthetic annotations (captions, VQA pairs) for existing images, creating multimodal datasets for fine-tuning.
- **Target Tasks:**
  - SemiKong is designed for process planning, optimization, and expert-level text reasoning in semiconductor manufacturing.
  - This project targets nanomaterial image understanding, enabling automated visual interpretation of SEM/TEM imagery.
- **Models:**
  - SemiKong builds on text-based LLMs (LLaMA-3 text), adapted for domain-specific language understanding.
  - This project adapts and fine-tunes open-source multimodal VLMs (Meta LLaMA-3.2 Vision, Qwen-VL, PaliGemma-2) for nanomaterial domains.