SemiKong: Curating, Training, and Evaluating a Semiconductor Industry-Specific Large Language Model

Christopher Nguyen[1], William Nguyen[1], Atsushi Suzuki[3], Daisuke Oku[3], Hong An Phan[1], Sang Dinh[1], Zooey Nguyen[1], Anh Ha[1], Vinh Luong[1], Shruti Raghavan[1], Huy Vo[2], Thang Nguyen[2], Lan Nguyen[2], Yoshikuni Hirayama[1]

[1]Aitomatic, Inc.
[2]FPT Software, AI Center
[3]Tokyo Electron Ltd

Abstract

Large Language Models (LLMs) have demonstrated the potential to address some issues within the semiconductor industry. However, they are often general-purpose models that lack the specialized knowledge needed to tackle the unique challenges of this sector, such as the intricate physics and chemistry of semiconductor devices and processes. SemiKong, the first industry-specific LLM for the semiconductor domain, provides a foundation that can be used to develop tailored proprietary models. With SemiKong 1.0, we aim to develop a foundational model capable of understanding etching problems at an expert level. Our key contributions include (a) curating a comprehensive corpus of semiconductor-related texts, (b) creating a foundational model with in-depth semiconductor knowledge, and (c) introducing a framework for integrating expert knowledge, thereby advancing the evaluation process of domain-specific AI models. Through fine-tuning a pre-trained LLM using our curated dataset, we have shown that SemiKong outperforms larger, general-purpose LLMs in various semiconductor manufacturing and design tasks. Our extensive experiments underscore the importance of developing domain-specific LLMs as a foundation for company- or tool-specific proprietary models, paving the way for further research and applications in the semiconductor domain. Code and dataset will be available at https://github.com/aitomatic/semikong[1].

Introduction

Semiconductor Manufacturing and Design

Semiconductors play an essential role in powering various electronic devices and driving development across industries such as telecommunications, automotive, healthcare, renewable energy, and IoT. In semiconductor manufacturing and design, the two main phases, FEOL and BEOL, each present their own unique challenges. FEOL, the front end of line processes, involves the creation of active devices on the semiconductor wafer. This includes steps such as wafer preparation, photolithography, etching, ion implantation, and gate oxide formation (El-Kareh 1994). These processes are crucial for defining the transistor structures and other active components of the integrated circuit (IC. On the other hand, BEOL, the back end of line processes, focuses on connecting the active devices created during FEOL. This includes the formation of metal layers, insulation, and bonding pads. Back-end processes are essential for establishing the electrical connections between devices and enabling the overall functionality of the IC (May and Spanos 2006). As feature sizes continue to shrink and device architectures become more complex, the need for advanced manufacturing techniques and design methodologies has become paramount. This has led to a growing interest in leveraging artificial intelligence (AI) and machine learning (ML) techniques to optimize semiconductor manufacturing processes and assist in design tasks (Amuru et al. 2022).

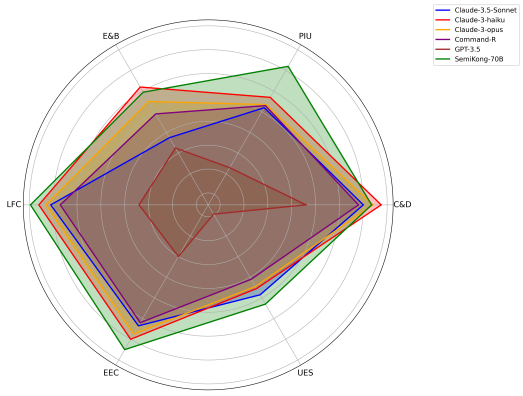

Figure 1: **Comparison of SemiKong and commercial models.** SemiKong is a open source foundation model but achieved comparable performance on E&B (Efficiency and Brevity), C&D (Clarity and Directness) with other commercial models and significantly outperformed these products in PIU (Practicality and Immediate Usability), LFC (Logical Flow and Coherence), EEC (Expert-to-Expert Communication), UES (Use of Examples and Specificity).

Use of LLMs in Semiconductors

Recent advancements in LLMs have demonstrated their remarkable potential in various domains, including the semiconductor industry (Liu et al. 2023a). LLMs, trained on vast amounts of text data using self-supervised learning techniques, have shown the ability to capture rich domain knowledge and generate human-like text. This has opened up new possibilities for applying LLMs to semiconductor process technology and IC design tasks. In the context of semiconductor process technology, LLMs can potentially assist in

[1]Open-Source AI for Mainstream Use Workshop, AAAI, 2025

tasks such as process parameter optimization, anomaly detection (Russell-Gilbert et al. 2024), and predictive maintenance of manufacturing equipment (Lee and Su 2023). By leveraging the vast amount of process data and domain knowledge embedded in the pre-trained models, LLMs can help identify patterns, predict process outcomes, and suggest optimal settings for various manufacturing steps. Similarly, in the realm of IC design, LLMs can aid in tasks such as design rule checking, layout generation, and design space exploration (Chang et al. 2023). By learning from large datasets of IC layouts and design rules (Mallappa et al. 2024), LLMs can potentially generate new designs that adhere to the specified constraints and optimize for desired performance metrics.

Furthermore, the open-source development paradigm is instrumental in enhancing the applicability and evolution of LLMs in specialized domains such as semiconductor manufacturing and IC design. Open-source frameworks enable the aggregation and dissemination of domain-specific knowledge, fostering collaboration among researchers, engineers, and industry practitioners. This collective effort accelerates innovation, reduces redundancy, and democratizes access to advanced tools and datasets, making cutting-edge technologies available to organizations of all sizes. In addition, the transparency provided by open source initiatives ensures reproducibility and reliability, which are critical to validate and benchmark LLM performance in real-world scenarios. For instance, open-source LLMs tailored to the semiconductor industry, such as SemiKong, can serve as a foundation for both academic research and proprietary model development, thereby bridging the gap between theoretical advancements and practical applications. This dual impact underscores the importance of open-source development as a key enabler for driving progress and achieving scalability in industry-specific AI solutions.

Purpose and Scope

Building on the success and potential of LLMs, this paper introduces SemiKong, the first industry-specific LLM tailored for the semiconductor domain, focusing on applications in semiconductor process technology and manufacturing. We aim to address the limitations of generic foundation models by curating a comprehensive semiconductor-related text corpus and developing a novel pre-training approach that leverages domain-specific knowledge. By doing so, we seek to demonstrate the potential of industry-specific LLMs in improving the performance of AI-driven solutions for semiconductor manufacturing tasks.

The scope of this work encompasses the following:

- The curation of a large-scale, semiconductor-specific text corpus focused on process technology and manufacturing
- The development of SemiKong, a foundation model, specifically focuses on the etching problems in the semiconductor industry
- The fine-tuning of SemiKong on industry-relevant data and tasks related to process optimization and control
- The introduction of a novel framework to leverage expert feedback in order to advance the LLMs-based evaluation approach for domain-specific AI models.
- The evaluation of SemiKong's performance compared to general-purpose LLMs
- The discussion of the implications and potential applications of industry-specific LLMs in semiconductor manufacturing

The main contributions of this paper are as follows:

- **SemiKong-Corpus:** We curate a comprehensive semiconductor-related text corpus, covering a wide range of topics related to semiconductor process technology and manufacturing. This corpus serves as the foundation for training SemiKong and captures the domain-specific knowledge essential for addressing manufacturing-related tasks.
- **SemiKong-Trainer:** We present SemiKong, a specialized foundation model with extensive knowledge of semiconductor manufacturing terminology and process flows, with a particular focus on etching. By pretraining and fine-tuning SemiKong with our carefully curated data, we have achieved substantial quality improvements in downstream tasks compared to generic LLMs and even commercial LLM-based products, as demonstrated in Figure 1.
- **SemiKong-Eval:** We develop a novel framework to effectively leverage expert's knowledge to advance the LLMs-based evaluation process and produce high-quality benchmarks. Besides, we conduct extensive evaluations to assess SemiKong's performance on industry-relevant benchmarks, such as process parameter optimization, anomaly detection, and predictive maintenance. Our results demonstrate SemiKong's superiority over general-purpose LLMs, highlighting the importance of developing industry-specific models for the semiconductor manufacturing domain.

The remainder of this paper is organized as follows: Section Related Works provides an overview of related work on the application of AI and LLM in the semiconductor industry. Section Semiconductor Ontology introduces the semiconductor ontology, with a focus on the front-end processes of semiconductor manufacturing. Section SemiKong: Semiconductor Industry Specific LLM outlines the methodology used to curate a semiconductor-specific text corpus and develop the pre-training approach. Section Experimental Result presents the experimental setup and results, comparing the performance of SemiKong with general-purpose LLMs across various text generation-based manufacturing tasks. Section Conclusion and Future Research Directions discusses the implications of the findings, potential future research directions, and concludes the paper.

Related Works

AI in semiconductor manufacturing

The application of Artificial Intelligence (AI) in semiconductor manufacturing has seen significant advancements, leveraging various AI methods to enhance the efficiency, yield, and quality of semiconductor fabrication processes. This section reviews the state-of-the-art AI approaches applied in different stages of semiconductor manufacturing, including two important steps: mask optimization, and hotspot

detection. Mask optimization is a critical step in semiconductor manufacturing. Traditional mask optimization methods typically consume significant runtime due to their iterative characteristics (Gu and Zakhor 2008). Recently, machine learning-based methods are proposed to accelerate mask optimization tasks (Choi, Shim, and Shin 2016). ILILT (Yang and Ren 2024) applied implicit learning for inverse lithography methods in mask optimization tasks. A large dataset, LithoBench (Zheng et al. 2023), consists of more than 120k circuit layout tiles for deep learning-based lithography simulation and mask optimization and is published to accelerate machine learning-based approaches. In addition, in the task of mask optimization, deep reinforcement learning is proposed to be applied to directly optimize the preferred objective in optical proximity correction (OPC) (Liang et al. 2024a). CAMO (Liang et al. 2024b) , a modulated reinforcement learning for correlation-aware mask optimization, is proposed to exploit the spatial correlation between the movements of neighboring segments.

Hotspot detection is an important step in semiconductor manufacturing to ensure the reliability and performance of integrated circuits (ICs). Hotspots are areas on a chip where excessive heat or stress can lead to defects, reducing yield and affecting the longevity and functionality of the devices. With the continuous scaling down of semiconductor technology nodes, the detection and mitigation of these hotspots have become increasingly significant. An active learning-based hotspot detection method (Yang et al. 2018) achieved an impressive performance in terms of detection accuracy. A new lithography hotspot detection framework based on the AdaBoost classifier and a simplified feature extraction (Matsunawa et al. 2015) obtained high accuracy with very low false alarms. In addition, semi-supervised learning with self-paced multi-task learning (Chen et al. 2019) is proposed for hotspot detection. Meanwhile, a hotspot detection using deep convolutional neural networks (Shin and Lee 2016) obtained accurate detection performance. These methods just focus on specific tasks rather than building a model to comprehensively support semiconductor operation engineers.

LLMs in the semiconductor industry

LLMs are proposed to adapt to domain-specific chip design, consisting of a wide range of tasks, from code generation to bug summarization and chatbot assistance for EDA engineers. ChipNemo (Liu et al. 2023b) developed by NVIDIA, proved that domain fine-tuned LLM models outperform general-purpose LLM models such as Llama3, and GPT4 in three specific tasks as engineering assistant chatbot for Q&A, EDA scripts generation, and bug summarization and analysis. RTLCoder (Liu et al. 2023c) outperforms GPT-3.5 in design RTL generation with an open-source dataset and a new training scheme via code quality feedback. ChipGPT (Chang et al. 2024) reinforces data-driven methods by making clear that data is all you need to fine-tune an LLM model for chip design, the results demonstrate a significant improvement in the code generation tasks with domain LLMs. Hdldebugger (Yao et al. 2024) focuses on using the LLM model for debugging via the LLM-assisted HDL debugging framework. Meanwhile, Rtlfixer (Tsai, Liu, and Ren 2023) targets fixing RTL syntax errors automatically with LLM models. Chip-Chat (Blocklove et al. 2023) conducted experiments with conversational LLMs to design and verify an 8-bit accumulator with GPT-4 and GPT-3.5. ChatEDA (He et al. 2023) introduces an autonomous agent for EDA empowered by a fine-tuned LLaMA2 70B model that outperforms the GPT-4 model in this task. In addition, inspired by LLMs in Natural Language Processing (NLP), Large Circuit Models (Chen et al. 2024) are proposed as a new paradigm to streamline the EDA process. However, these models are mostly developed with small public datasets with the limitation of the expert's participation in the development process.

LLM as an evaluator

Human evaluation is a crucial method for assessing Natural Language Generation (NLG) algorithms (Guzmán et al. 2015). Many NLP tasks require skilled annotators or experts for reliable evaluations (Gillick and Liu 2010). However, recruiting human experts is often impractical due to high costs and concerns about reproducibility. Meanwhile, automatic metrics like BLEU (Papineni et al. 2002) and ROUGE (Lin 2004) fall short of reliability expectations and fail to reflect human preferences accurately. Recently, using LLMs to evaluate NLG (Wang et al. 2023) has been introduced to address these issues. These methods are reference-free, asking LLMs to justify their answers based on task requirements and demonstrating correlation with human judgment, assuming that LLMs can understand and assign higher probabilities to high-quality and fluent texts. G-eval (Liu et al. 2023d) applied the chain-of-thought technique by asking LLMs to generate detailed evaluation steps to enhance evaluation quality. Despite these advancements, these methods share a common limitation: they assume LLMs can inherently understand and evaluate knowledge. However, experts with many years of experience are often needed to evaluate complex questions in domains requiring deep expertise, such as semiconductors, for accurate judgments. Given these challenges, this paper proposes a framework that leverages expert feedback to create criteria for more reliable assessments by LLMs, approaching expert-level reliability. This feedback is also used to generate a high-quality benchmark for the semiconductor domain. OSCaR (Nguyen et al. 2024) employed a similar approach in generating high-quality benchmarks. However, they utilized feedback from normal humans on Amazon MTurk, while our benchmark relies on expert knowledge, ensuring significantly higher reliability.

Semiconductor Ontology

Semiconductor manufacturing involves numerous complex steps and processes, requiring extensive knowledge for effective execution. In each step, having an expert specialized in that particular field to guide workers is crucial. However, the semiconductor manufacturing process is not easily accessible to AI researchers, who possess deep expertise in AI but often lack domain-specific knowledge, particularly an understanding of semiconductor manufacturing. This gap hinders the development of efficient, domain-specific AI models. To address this challenge, we collaborated with

semiconductor experts to develop an ontology that systematically structures the entire semiconductor manufacturing process. This ontology is constructed using a top-down approach, dividing the field from general to detailed levels, sub-levels, and specific processes, ensuring that no critical process is overlooked.

By systematically structuring the semiconductor manufacturing process, our ontology not only addresses the knowledge gap for AI researchers but also serves as a foundation for creating more effective domain-specific AI models. This ontology is invaluable not only for building specialized AI models, like SemiKong for etching, but also serves as a benchmark for evaluating future general intelligence models that aim to address a wide range of semiconductor manufacturing topics, both in model development and evaluation. The hierarchical structure of the ontology enhances understanding and training efficiency, enabling the creation of specialized language model agents with precise insights tailored to specific stages of semiconductor manufacturing. Consequently, this ontology serves as a dynamic tool for guiding future training efforts and ensuring that language models remain up-to-date with industry advancements. To achieve these objectives, a well-designed procedure and meticulous implementation are essential in constructing a comprehensive semiconductor ontology.

Our ontology for semiconductor manufacturing was developed in collaboration with industry experts to cover the entire semiconductor manufacturing process, from front-end to back-end, including Substrate Preparation, Film Formation, Patterning, Doping, Planarization, Cleaning and Surface Preparation, Thermal Processing, Metrology and Inspection, Advanced Modules, and Back-End Processes. These represent the primary levels of semiconductor fabrication, which our experts further divided into secondary and tertiary levels. For example, Patterning is a key first-level process, which is further broken down in the second level into subclasses such as Etching. The third level categorizes Etching into Wet Etching, Dry Etching, Plasma Etching, Reactive Ion Etching, Deep Reactive Ion Etching, Isotropic Wet Etching, Anisotropic Wet Etching, Atomic Ion Etching, and Electron Cyclotron Etching. This paper introduces our model, SemiKong, which can comprehensively understand and provide support for the etching process, ensuring that our ontology fully covers this critical area and lays the groundwork for future specialized models in other semiconductor manufacturing processes. In this work, we mostly focused on the etching process. This highlights that there is still significant scope for further research in this field.

Semiconductor Industry Specific LLM

Developing an expert-level, domain-specific model necessitates acquiring in-depth knowledge in the relevant field. A prevalent approach involves training models with comprehensive domain-specific data. This training process can be divided into two stages: pretraining and fine-tuning. Although this method typically leads to significant model improvements, it still presents challenges related to data quality assurance, defining the model training strategy, and determining appropriate evaluation metrics. In this section, we will discuss our data curation pipeline (Section Data Curation), the process for training the SemiKong model using both pretraining and fine-tuning (Section Model tranining), and the incorporation of expert feedback in the evaluation pipeline (Section The proposed method for evaluating LLM in semiconductor manufacturing).

Data Curation

High-quality domain-specific datasets, including those for the semiconductor domain, are often scarce. To address this problem, we introduce a large-scale, high-quality text-based dataset specifically for the semiconductor domain. Our dataset consists of two parts: documents for pretraining and instructions for fine-tuning.

Pre-Training dataset: Pretraining is a crucial step for incorporating knowledge into models. However, pretrained generic models often prioritize data coverage over depth. It is challenging to determine which data was used to train the model and the extent of the knowledge it encompasses. Based on this issue, we assume that generic pretrained models lack in-depth knowledge and the ability to focus on specific domains. We introduce a text-based dataset focused on semiconductors, extracted from technical books, papers, and patents. To construct this dataset, we automatically crawled for public PDF documents available on the internet, including publications from arxiv and semiconductors-related open books. These documents were then converted to raw text using the PyPDF library. Since the raw text often has formatting issues, we employed GPT-4o-mini for post-processing to transform the text into markdown format. This step not only corrected parsing errors but also preserved special types of information, such as tables. The effectiveness of our proposed pretraining data set is demonstrated in the experimental results shown in Table 4. The results indicate significant improvement when comparing a model purely fine-tuned with instructions to a model pre-trained with our dataset before fine-tuning.

Instruction dataset: We utilized GPT-4o and GPT-o1-preview to generate semiconductor-related instructions. Starting with a predefined list of semiconductor terms, GPT-4o expanded this into a broader set of synonyms and related keywords. Using this enriched list, GPT-4o generated 50,000 questions, categorized as follows: 5,000 on semiconductor concepts, 5,000 on complex etching problems requiring mathematical reasoning, and 40,000 on standard etching process issues (Table 1). GPT-4o provided answers for conceptual and routine questions, while GPT-o1-preview addressed math-intensive and complex reasoning problems. This approach enhances the model's ability to tackle semiconductor etching challenges, making it a robust foundation in the domain.

Table 1: Details of SemiKong dataset

Dataset Details	**Quantity**
Books, book chapters	129
Etching research papers	708
Research papers	20K
Instructions	50K
Tokens	525.6M

Model tranining

The curated dataset described in Section 3.1 was employed to train our SemiKong models. Initially, the text data was tokenized using Tiktoken, a tokenizer based on BPE, which is widely utilized in numerous NLP applications. Subsequently, Rotary Position Embedding (RoPE) was incorporated as the positional embedding component to enable the LLM to capture positional information effectively. The training process comprised two stages: model pre-training using our pure text dataset and supervised fine-tuning (SFT). Then, we do post-training processing to make models become more suitable for production. The model overview and computational resources are detailed in Table 2.

Model pre-training: We hypothesized that generic pre-trained models lack domain-specific knowledge. Therefore, we pre-trained our SemiKong models using the Llama3 8B and 70B checkpoints from Meta as a starting point. This step aims to enhance the models' in-depth semiconductor domain-specific knowledge, thereby ensuring that they focus more on the specific domain in which we intend for the models to serve as experts in the future.

Supervised fine-tuning (SFT): While pre-training endows the models with extensive domain knowledge, fine-tuning empowers them to effectively handle anticipated tasks such as question-answering, dialogue, and reasoning. Leveraging the availability of instruction data, we employ SFT to guide the models in addressing semiconductor-related tasks. We have adopted LoRA for SFT, and our results indicate that transitioning to full fine-tuning could potentially enhance performance even further.

Post-training process: Following pre-training and fine-tuning, we conducted quantization and merging to prepare the models for deployment. Our implementation utilized GPTQ [30], an accurate post-training quantization technique for generative pre-trained transformers. Finally, the LoRA adapter was merged with the original LLM model to produce the final LLM model tailored for semiconductor manufacturing.

The proposed method for evaluating LLM in semiconductor manufacturing

The evaluation of AI assistant models in domain-specific contexts requires expert judgment to justify the usefulness of the model's responses. However, expert annotations are often limited and costly. Therefore, developing an automated metric to assess the quality of these models is crucial for their development and evaluation. Such a metric not only supports project development but also serves as a standard for future research in this area. Motivated by this need, we propose a novel pipeline to generate a list of criteria for evaluation. This list of criteria will be fed into LLMs to enhance their ability to justify expert models. A key challenge is that different subfields require different evaluation criteria, and no universal criteria apply to all problems. We anticipate that with the finalized list of criteria, LLMs will be able to evaluate the responses from AI assistant models with a high correlation to expert judgment. Our contribution includes developing a pipeline for generating a customized list of criteria by leveraging expert feedback. We demonstrate the effectiveness of our pipeline by generating a list of criteria for the semiconductor industry domain. It is important to emphasize that our method is not only applicable to the semiconductor domain but also to other domains requiring human expertise. In our proposed pipeline for evaluation, we initially collected a set of questions from three primary sources: 737 questions from our company's experts, 150 questions crawled from the ResearchGate forum, and 100 general questions generated by ChatGPT. Our internal experts carefully reviewed and evaluated each question to ensure its quality. Following this review, questions were classified into three difficulty levels: Easy, Medium, and Hard, as detailed in Table 3. Additionally, our experts developed an ontology, as detailed in Section , to categorize the questions' processes into high, sub, and specific levels. Finally, we utilized all the collected questions and annotations, inputting them into GPT-4o and our SemiKong model to generate the initial answers.

Building upon the human-in-the-loop concept, we have advanced it to an expert-in-the-loop framework. As shown in Figure II, in this approach, experts review the initial answers generated by LLMs. These experts, who possess extensive knowledge in their fields, not only provide correct answers but also evaluate the quality of other answers. This dual capability allows us to generate ground truth for benchmarking and to synthesize a set of criteria to guide LLMs in evaluating semiconductor expert models. To implement this, we request experts to score the answers and provide detailed justifications for their scores. Machine learning researchers then analyze these justifications to develop a comprehensive criteria list, which is used to guide LLMs to score model outputs. The goal is to create clear, precise criteria that enable LLMs to make evaluations similar to those of human experts. This process is iterative, with continuous updates to the criteria based on new data annotations from experts, thereby progressively improving the evaluation framework. In this paper, we define the criteria for using LLMs to evaluate semiconductor expert models as follows:

Clarity and Directness (C&D): This criterion involves using simple and straightforward language to ensure that the answer is easily understood. This means avoiding unnecessary jargon or technical terms that could confuse the reader. It also requires directly addressing the question or topic at hand in each sentence, maintaining a focus on the main points. Organizing information with bullet points or numbered lists can further enhance readability and make key points more accessible.

Practicality and Immediate Usability (PIU): Practicality and immediate usability involve providing recommendations that are both practical and easy to implement. This means focusing on clear, actionable steps rather than theoretical explanations, ensuring that the guidance is directly applicable to real-world situations. Recommendations should be realistic and suitable to the specific context, making them immediately usable and relevant to the audience's needs.

Efficiency and Brevity (E&B): Efficiency and brevity involve eliminating redundant information and combining related points to avoid verbosity. The goal is to keep the in-

Table 2: Overview of models

Feature	SemiKong-7B	SemiKong-80B
Number of trainable parameters	2,849,712	50,732,193
Hardware Resources for Training	NVIDIA GPU 4x A100 80GB	NVIDIA GPU 8x A100 80GB
Training time	∼150 hours (15 runs)	∼200 hours (2 runs)

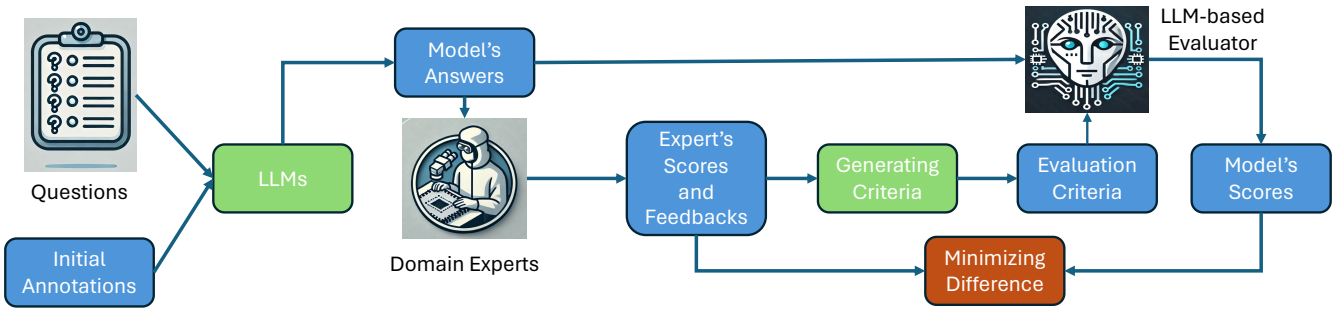

Figure 2: **The evaluation benchmark development pipeline.**

Table 3: Difficulty levels of questions in the evaluation dataset

Category	#Question
Easy	100
Medium	737
Hard	150

formation concise while still covering all necessary details, ensuring that the message is clear and to the point without unnecessary elaboration.

Logical Flow and Coherence (LFC): Logical flow and coherence involve arranging points in a clear, logical order to make the answer easy to follow. This includes grouping related points together under clear categories, enhancing the overall coherence, and ensuring that the user can easily understand the progression of ideas.

Expert-to-Expert Communication (EEC): Expert-to-expert communication involves tailoring responses as instructions or directions that an experienced engineer would give to another engineer in the same role but with less experience. This ensures that the conversation is part of a problem-solving process, focusing on advanced concepts and practical guidance without delving into overly basic explanations that would be unnecessary for an expert audience.

Use of Examples and Specificity (UES): The use of examples and specificity involves providing examples only when they add significant value to the explanation. Ensure that comparisons are directly related to the point being made and are concise. Introduce technical terms only if they are essential to the discussion, and offer concise explanations for these terms only if requested to maintain clarity and relevance.

Experimental Result

Implementation details

For training Semikong, we utilized 8 NVIDIA A100 80GB GPUs. We followed guidelines from Transformers HuggingFace, HuggingFace Accelerator, and the LLaMA-Factory library for fine-tuning a LLM. The hyperparameters for pre-training and SFT included a batch size of 3, gradient accumulation steps of 3, and a learning rate of 1.0e-5. The training targeted 50 million (0.05B) parameters and was conducted over 5 epochs with QLoRA, employing a cosine learning rate scheduler with a warm-up ratio of 0.15. NF4 quantization was used for efficient parameter representation. Additionally, 20% of the dataset was allocated for validation.

Evaluation

To evaluate the contribution of finetuning and pre-training, we conducted experiments to compare three models: Llama3, SemiKong SFT only, and pretrained SemiKong with SFT. Table 4 shows the result of our experiment. In general, fine tuning only did not improve the performance of the model. It shows that generic models lack knowledge of domain specifics. When the model is pre-trained to learn more in-depth knowledge, the model's performance begins to show signs of improvement. However, the model implemented for this experiment only has 8B parameters, which limits the ability to learn knowledge of the model. So, in the next experiments, we will conduct the experiment on a larger model with 70B parameters and fine-tune only the model that has been pre-trained with our proposed semiconductor dataset.

The experimental results in Table 5 demonstrate that models with 70B parameters significantly outperform those with 8B parameters. Even when compared to our fine-tuned SemiKong 8B model, the base Llama3 70B model still outperforms it. Based on this observation, our SemiKong 70B model and the experimental results show that our approach significantly surpasses both the generic open-source Llama3 8B and Llama3 70B models across all criteria.

To demonstrate the superiority of SemiKong, we conducted experiments comparing its performance with that of commercial products. It's important to note that SemiKong

Table 4: Comparison of contribution of SFT and pre-training

Model	C&D	PIU	E&B	LFC	EEC	UES	Total
Llama3 8B	3.65	3.35	3.07	3.67	3.47	**3.28**	20.49
SemiKong 8B (SFT only)	3.61	**3.36**	3.22	3.64	3.52	3.16	20.51
SemiKong 8B (Pretraining+SFT)	**3.73**	3.35	**3.40**	**3.68**	**3.54**	3.11	**20.81**

Table 5: Compare with open source models

Model	C&D	PIU	E&B	LFC	EEC	UES	Total
Llama3 8B	3.65	3.35	3.07	3.67	3.47	3.28	20.49
SemiKong 8B	3.73	3.35	3.40	3.68	3.54	3.11	20.81
Llama3 70B	3.89	3.63	3.55	3.99	3.82	3.47	22.35
SemiKong 70B (Pretraining+SFT)	**4.07**	**4.05**	**3.88**	**4.23**	**4.13**	**3.66**	**24.02**

Table 6: Compare with commercial products

Model	C&D	PIU	E&B	LFC	EEC	UES	Total
Claude-3.5-Sonnet	3.80	3.44	3.15	3.82	3.67	3.37	21.25
Claude-3-haiku	**3.95**	3.54	**3.64**	3.92	3.80	3.31	22.16
Claude-3-opus	3.88	3.47	3.50	3.86	3.75	3.29	21.75
Command-R	3.76	3.46	3.38	3.74	3.64	3.22	21.20
GPT-3.5	3.32	2.86	3.05	3.08	3.00	2.59	17.90
SemiKong-70B	3.87	**3.84**	3.59	**3.99**	**3.90**	**3.46**	**22.65**

is a foundation model and does not rely on supporting systems like RAG. As shown in Table 6 and Figure 1, SemiKong delivers comparable performance in the C&D and E&B metrics, while it outperforms in four out of six key metrics: PIU, LFC, EEC, and UES. These metrics are critical for determining whether a model meets the needs of an expert. Overall, SemiKong achieves state-of-the-art performance, making it the most suitable model for expert use. Its practicality for immediate application, logical flow, avoidance of unnecessary information, and ability to provide concise and accurate answers are exactly what engineers require for their daily work.

Conclusion and Future Research Directions

In this paper, we introduce SemiKong, the first foundation model specialized for the semiconductor industry, available in both 8B and 70B versions. Additionally, we have publicized a large-scale dataset tailored for semiconductor applications, encompassing both pretraining and fine-tuning data. We also present a semiconductor ontology designed to support AI researchers in developing new AI research within the semiconductor field. Our SemiKong models have achieved state-of-the-art performance, outperforming open-source foundation models and surpassing commercial products in expert use. However, SemiKong represents just the initial effort, and there remains significant work to be done. First, based on our proposed ontology, we can further develop additional processes beyond etching, making AI for semiconductors more comprehensive and applicable to various stages of semiconductor manufacturing. Secondly, our pipeline can be adapted and expanded to other industries, thereby enhancing operations across multiple sectors.

Acknowledgments

We would like to express our gratitude to the AI Alliance (https://thealliance.ai) for providing the impetus, resources, and platform for this work, and for collaboration in open science. We also extend our thanks to the member organizations of the AI Alliance, their researchers, and engineers for their valuable contributions to this study, including Anthony Annunziata (IBM Research), Sean Hughes (ServiceNow), Phong Nguyen (FPT Software, AI Center), Noritaka Yokomori (Tokyo Electron). Their expertise, insights, and collaborative spirit have been instrumental in advancing our research.

References

Amuru, D.; Vudumula, H. V.; Cherupally, P. K.; Gurram, S. R.; Ahmad, A.; Zahra, A.; and Abbas, Z. 2022. AI/ML Algorithms and Applications in VLSI Design and Technology. *Integr.*, 93: 102048.

Blocklove, J.; Garg, S.; Karri, R.; and Pearce, H. A. 2023. Chip-Chat: Challenges and Opportunities in Conversational Hardware Design. *2023 ACM/IEEE 5th Workshop on Machine Learning for CAD (MLCAD)*, 1–6.

Chang, K.; Wang, K.; Yang, N.; Wang, Y.; Jin, D.; Zhu, W.; Chen, Z.; Li, C.; Yan, H.; Zhou, Y.; Zhao, Z.; Cheng, Y.; Pan, Y.; Liu, Y.; Wang, M.; Liang, S.; Han, Y.; Li, H.; and Li, X. 2024. Data is all you need: Finetuning LLMs for Chip Design via an Automated design-data augmentation framework. *ArXiv*, abs/2403.11202.

Chang, K.; Wang, Y.; Ren, H.; Wang, M.; Liang, S.; Han, Y.; Li, H.; and Li, X. 2023. ChipGPT: How far are we from natural language hardware design. *ArXiv*, abs/2305.14019.

Chen, L.; Chen, Y.; Chu, Z.; Fang, W.; Ho, T.-Y.; Huang, Y.; Khan, S.; Li, M.; Li, X.; Liang, Y.; Lin, Y.; Liu, J.; Liu, Y.;

Luo, G.; Shi, Z.; Sun, G.; Tsaras, D.; Wang, R.; Wang, Z.; Wei, X.; Xie, Z.; Xu, Q.; Xue, C.; Young, E. F. Y.; Yu, B.; jie Yuan, M.; Zhang, H.; Zhang, Z.; Zhao, Y.; Zhen, H.-L.; Zheng, Z.; Zhu, B.; Zhu, K.; and Zou, S. 2024. The Dawn of AI-Native EDA: Promises and Challenges of Large Circuit Models. *ArXiv*, abs/2403.07257.

Chen, Y.; Lin, Y.; Gai, T.; Su, Y.; Wei, Y.; and Pan, D. Z. 2019. Semi-Supervised Hotspot Detection with Self-Paced Multi-Task Learning. *2019 24th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 1–6.

Choi, S.; Shim, S.; and Shin, Y. 2016. Machine learning (ML)-guided OPC using basis functions of polar Fourier transform. In *Advanced Lithography*.

El-Kareh, B. 1994. Fundamentals of Semiconductor Processing Technology.

Gillick, D.; and Liu, Y. 2010. Non-Expert Evaluation of Summarization Systems is Risky. In *Mturk@HLT-NAACL*.

Gu, A.; and Zakhor, A. 2008. Optical Proximity Correction With Linear Regression. *IEEE Transactions on Semiconductor Manufacturing*, 21: 263–271.

Guzmán, F.; Abdelali, A.; Temnikova, I.; Sajjad, H.; and Vogel, S. 2015. How do Humans Evaluate Machine Translation. In *WMT@EMNLP*.

He, Z.; Wu, H.; Zhang, X.; Yao, X.; Zheng, S.; Zheng, H.; and Yu, B. 2023. ChatEDA: A Large Language Model Powered Autonomous Agent for EDA. *2023 ACM/IEEE 5th Workshop on Machine Learning for CAD (MLCAD)*, 1–6.

Lee, J.; and Su, H. 2023. A unified industrial large knowledge model framework in Industry 4.0 and smart manufacturing. *International Journal of AI for Materials and Design*.

Liang, X.; Ouyang, Y.; Yang, H.; Yu, B.; and Ma, Y. 2024a. RL-OPC: Mask Optimization With Deep Reinforcement Learning. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43: 340–351.

Liang, X.; Yang, H.; Liu, K.; Yu, B.; and Ma, Y. 2024b. CAMO: Correlation-Aware Mask Optimization with Modulated Reinforcement Learning. *ArXiv*, abs/2404.00980.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Liu, M.; Ene, T.-D.; Kirby, R.; Cheng, C.; Pinckney, N.; Liang, R.; Alben, J.; Anand, H.; Banerjee, S.; Bayraktaroglu, I.; Bhaskaran, B.; Catanzaro, B.; Chaudhuri, A.; Clay, S.; Dally, B.; Dang, L.; Deshpande, P.; Dhodhi, S.; Halepete, S.; Hill, E.; Hu, J.; Jain, S.; Khailany, B.; Kunal, K.; Li, X.; Liu, H.; Oberman, S. F.; Omar, S.; Pratty, S.; Raiman, J.; Sarkar, A.; Shao, Z.; Sun, H.; Suthar, P. P.; Tej, V.; Xu, K.; and Ren, H. 2023a. ChipNeMo: Domain-Adapted LLMs for Chip Design. *ArXiv*, abs/2311.00176.

Liu, M.; Ene, T.-D.; Kirby, R.; Cheng, C.; Pinckney, N.; Liang, R.; Alben, J.; Anand, H.; Banerjee, S.; Bayraktaroglu, I.; Bhaskaran, B.; Catanzaro, B.; Chaudhuri, A.; Clay, S.; Dally, B.; Dang, L.; Deshpande, P.; Dhodhi, S.; Halepete, S.; Hill, E.; Hu, J.; Jain, S.; Khailany, B.; Kunal, K.; Li, X.; Liu, H.; Oberman, S. F.; Omar, S.; Pratty, S.; Raiman, J.; Sarkar, A.; Shao, Z.; Sun, H.; Suthar, P. P.; Tej, V.; Xu, K.; and Ren, H. 2023b. ChipNeMo: Domain-Adapted LLMs for Chip Design. *ArXiv*, abs/2311.00176.

Liu, S.; Fang, W.; Lu, Y.; Zhang, Q.; Zhang, H.; and Xie, Z. 2023c. RTLCoder: Outperforming GPT-3.5 in Design RTL Generation with Our Open-Source Dataset and Lightweight Solution. *ArXiv*, abs/2312.08617.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023d. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Conference on Empirical Methods in Natural Language Processing*.

Mallappa, U.; Mostafa, H.; Galkin, M.; Phielipp, M.; and Majumdar, S. 2024. FloorSet - a VLSI Floorplanning Dataset with Design Constraints of Real-World SoCs. *ArXiv*, abs/2405.05480.

Matsunawa, T.; Gao, J.-R.; Yu, B.; and Pan, D. Z. 2015. A new lithography hotspot detection framework based on AdaBoost classifier and simplified feature extraction. In *Advanced Lithography*.

May, G. S.; and Spanos, C. J. 2006. Fundamentals of Semiconductor Manufacturing and Process Control.

Nguyen, N.; Bi, J.; Vosoughi, A.; Tian, Y.; Fazli, P.; and Xu, C. 2024. OSCaR: Object State Captioning and State Change Representation. *ArXiv*, abs/2402.17128.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.

Russell-Gilbert, A.; Sommers, A.; Thompson, A.; Cummins, L.; Mittal, S.; Rahimi, S.; Seale, M.; Jaboure, J.; Arnold, T.; and Church, J. 2024. AAD-LLM: Adaptive Anomaly Detection Using Large Language Models.

Shin, M.; and Lee, J.-H. 2016. Accurate lithography hotspot detection using deep convolutional neural networks. *Journal of Micro/Nanolithography, MEMS, and MOEMS*, 15.

Tsai, Y.-D.; Liu, M.; and Ren, H. 2023. RTLFixer: Automatically Fixing RTL Syntax Errors with Large Language Models. *ArXiv*, abs/2311.16543.

Wang, J.; Liang, Y.; Meng, F.; Sun, Z.; Shi, H.; Li, Z.; Xu, J.; Qu, J.; and Zhou, J. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. *ArXiv*, abs/2303.04048.

Yang, H.; Li, S.; Tabery, C.; Lin, B.; and Yu, B. 2018. Bridging the Gap Between Layout Pattern Sampling and Hotspot Detection via Batch Active Learning. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40: 1464–1475.

Yang, H.; and Ren, H. 2024. ILILT: Implicit Learning of Inverse Lithography Technologies. *ArXiv*, abs/2405.03574.

Yao, X.; Li, H.; Chan, T. H.; Xiao, W.; Yuan, M.; Huang, Y.; Chen, L.; and Yu, B. 2024. HDLdebugger: Streamlining HDL debugging with Large Language Models. *ArXiv*, abs/2403.11671.

Zheng, S.; Yang, H.; Zhu, B.; Yu, B.; and Wong, M. D. F. 2023. LithoBench: Benchmarking AI Computational Lithography for Semiconductor Manufacturing. In *Neural Information Processing Systems*.