

ChatEDA: A Large Language Model Powered Autonomous Agent for EDA

Zhuolun He^{1,2†}, Haoyuan Wu^{2†}, Xinyun Zhang¹, Xufeng Yao¹, Su Zheng¹, Haisheng Zheng², Bei Yu¹
¹Chinese University of Hong Kong ²Shanghai Artificial Intelligent Laboratory

Abstract—The integration of a complex set of Electronic Design Automation (EDA) tools to enhance interoperability is a critical concern for circuit designers. Recent advancements in large language models (LLMs) have showcased their exceptional capabilities in natural language processing and comprehension, offering a novel approach to interfacing with EDA tools. This research paper introduces ChatEDA, an autonomous agent for EDA empowered by a large language model, AutoMage, complemented by EDA tools serving as executors. ChatEDA streamlines the design flow from the Register-Transfer Level (RTL) to the Graphic Data System Version II (GDSII) by effectively managing task planning, script generation, and task execution. Through comprehensive experimental evaluations, ChatEDA has demonstrated its proficiency in handling diverse requirements, and our fine-tuned AutoMage model has exhibited superior performance compared to GPT-4 and other similar LLMs.

I. INTRODUCTION

Electronic Design Automation (EDA) encompasses a crucial set of software tools utilized for circuit design, analysis, and verification. These tools are organized within a complex design flow, featuring intricate programming interfaces. Notably, advanced RTL-to-GDSII design platforms like OpenROAD [1] and iEDA [2] consist of numerous procedures and adjustable parameters. Commercial tools, with their extensive functionalities and options, offer even more comprehensive capabilities. In practice, circuit design engineers employ these tools iteratively to achieve their design objectives, often resorting to custom scripts for specific operations. Conventionally, scripting languages such as TCL have been the *de facto* means of interacting with EDA tools [3], which is tedious and prone to errors.

Recently, the field of natural language processing (NLP) has undergone a revolutionary transformation with the emergence of large language models (LLMs) such as GPT-3 [4], GPT-4 [5], Claude2 [6], and Llama [7], [8]. Extensive training on large corpora enables large language models to acquire emergent abilities [9], including the capacity to reason and comprehend complex concepts, by learning intricate patterns and relationships in language. This allows these models to demonstrate remarkable accuracy and fluency in a variety of NLP tasks, such as natural language understanding and generation. To leverage their potential in specialized domains [10]–[12], instruction tuning [13] fine-tunes LLMs with domain-specific corpora, resulting in remarkable performance on these specialized domains. Specifically, Vicuna [14], Guanaco [15], and Orca [16], have applied instruction tuning to train LLMs,

[†]Equal Contributors

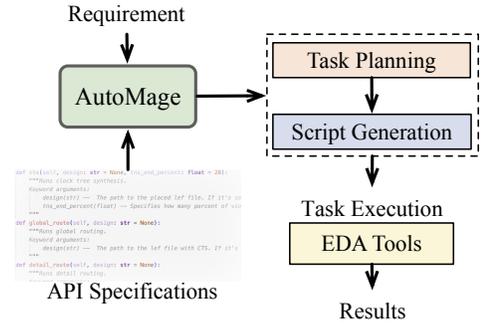


Fig. 1 Overview of AutoMage powered ChatEDA. With AutoMage as the controller and EDA tools as the executors, the workflow consists of three stages: 1) Task planning; 2) Script Generation; 3) Task execution;

making use of the outputs produced by the GPT, and thereby achieving significant outcomes.

Furthermore, scholars have initiated exploration into the incorporation of tools or models into LLMs. Toolformer [17], a groundbreaking methodology, integrates external API tags into text sequences, thus facilitating LLMs to connect with external tools. This tool utilization, coupled with the capacity for logical reasoning, broadens the LLM’s potential as a robust general problem solver. Several proof-of-concept demonstrations, including AutoGPT [18] and BabyAGI [19] serve as motivational illustrations.

In this context, we propose ChatEDA, an expert LLM designed to generate code for manipulating EDA tools based on natural language instructions. In other words, ChatEDA is an LLM-driven autonomous agent system for EDA, functioning as the agent’s intellectual hub, responding to human instructions and manipulating the EDA tools via APIs to deliver autonomous Register-Transfer Level (RTL) to Graphic Data System Version II (GDSII) capabilities without necessitating any code writing. Our contributions include 1) ChatEDA, the first LLM-powered EDA interfacing framework and methodology; 2) a fine-tuned language model, AutoMage, that is purpose-built to enhance the capabilities of ChatEDA; and 3) comprehensive evaluations to show the superior performance of ChatEDA and AutoMage, surpassing GPT-4 and other well-known LLMs in various tasks.

II. THE SCOPE OF CHATEDA

ChatEDA serves as an autonomous agent tailored for Electronic Design Automation (EDA) tools, providing users with a conversational interface to interact with these tools. Specifically, users can express their requirements in natural language prompts,

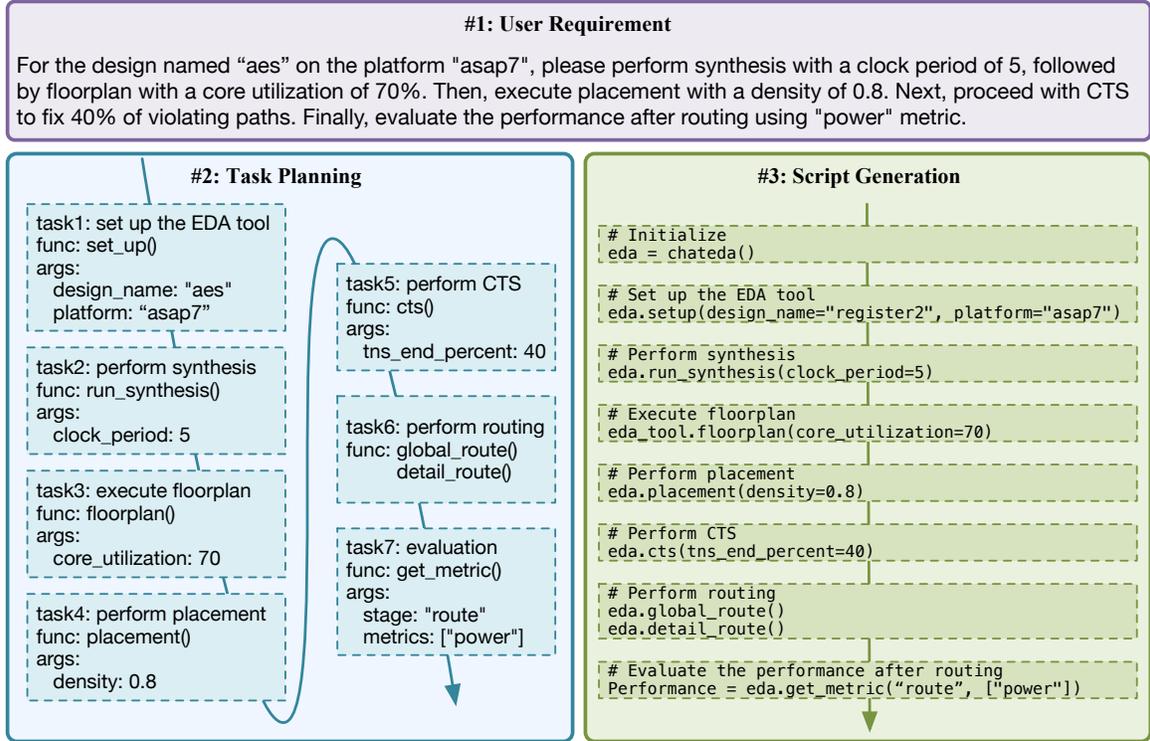


Fig. 2 Language functions as a conduit enabling ChatEDA to integrate EDA tools for resolving complex EDA tasks. Within the framework, ChatEDA acts as the controller that harmonizes and orchestrates the collaboration among various tools. ChatEDA first formulates a task list derived from user requirements, subsequently generating scripts corresponding to these decomposed tasks.

and ChatEDA’s objective is to generate executable programs (scripts) that align with the user’s specific needs.

Problem 1. (Target of ChatEDA) Develop an LLM-driven framework capable of processing natural language command as input, and generates an executable program to fulfill user requirements through the invocation of EDA tool APIs.

III. CHATEDA

ChatEDA, an LLM powered agent, is specifically designed for RTL-to-GDSII flow automation. The main objective of ChatEDA is to understand and respond to user requirements in natural language. In order to achieve this, ChatEDA is capable of breaking down complex user requirements into smaller, more manageable sub-tasks and subsequently utilizing appropriate EDA tools to address them.

As illustrated in Fig. 1, AutoMage, an LLM fine-tuned with EDA expert knowledge, serves as the central processing unit of ChatEDA. After receiving a natural language requirement from the user, AutoMage first interprets the requirement and decomposes it into a set of sub-tasks, known as task planning. Then, based on the decomposed smaller tasks and the specifications for the external tools, e.g., OpenROAD, AutoMage generates python scripts for accomplishing these tasks. Ultimately, ChatEDA executes the generated script to get the final output for the user requirement. We will detail the workflow and the training process of AutoMage in the following sections.

A. Task Planning

Within the scope of automating the RTL-to-GDSII flow via EDA tools, a myriad of user requests often encapsulate complex intentions. Primarily, it is crucial for the agent to understand human natural language requirements. Owing to the robust performance of AutoMage, ChatEDA can interpret the tasks that need to be executed based on human requirements.

Furthermore, given the complexity of automating the RTL-to-GDSII flow, orchestrating several sub-tasks becomes necessary to achieve the anticipated result. Hence, we propose task planning as the first stage of ChatEDA. This phase employs AutoMage to assess the user requirement and decompose it into a series of structured tasks subsequently. Fig. 2 illustrates an example of the task planning process. For the given requirement expressed in long and complicated natural language, ChatEDA can decompose this into a series of ordered sub-tasks via AutoMage, which can be easily handled by leveraging the EDA tools, such as logic synthesis, floorplan, placement, *etc.*

B. Script Generation

After the completion of the task planning phase, manageable sub-tasks are devised, thereby promoting the efficient orchestration of the complex task. Each sub-task is executable through the corresponding APIs from the EDA tools. Subsequently, it becomes necessary to formulate a script to invoke these APIs for task execution. During the script generation phase as depicted in Fig. 2, a sequence of text encompassing the API specifications, user requirements, and the decomposed sub-tasks, is utilized as the input for AutoMage. Then, AutoMage will outcome a Python script that can be executed directly

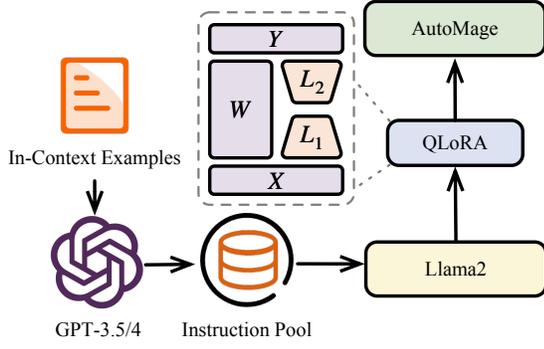


Fig. 3 Overview of Instruction Tuning. During the instruction tuning process, we use the self-instruction paradigm to construct our instruction pool via GPT. Then we apply the QLoRA technique for efficient instruction fine-tuning.

to perform an RTL-to-GDSII flow for efficient architectural and design space exploration, early Quality of Results (QoR) prediction, and detailed physical design implementation.

C. Task Execution

After script generation, ChatEDA runs the script with the Python interpreter, and subtasks are then executed by leveraging EDA tools. Given its demonstrated efficacy in script generation and task execution, ChatEDA can be a critical agent for reliable RTL-to-GDSII flow automation.

D. Instruction Tuning for AutoMage

To guarantee the reliability of ChatEDA, knowing when and how to use the tools, which is determined by the LLM’s capability, is of vital importance. AutoMage is an expert LLM that specializes in the utilization of EDA tools, which is barely learned in open-source LLMs. To integrate the expert knowledge into LLMs, instruction tuning is an effective approach that enables LLMs to benefit from the pairing of domain-specific natural language descriptions and their corresponding responses. Therefore, we incorporate instruction tuning to train AutoMage, the core controller of ChatEDA, based on the open-source LLMs (Llama2 [8]). The process of instruction tuning mainly includes three parts, self instruction, instruction collection, and instruction fine-tuning, which detailed as following.

Self Instruction. For instruction tuning, we need to collect high-quality instructions to teach LLMs how to utilize EDA tools via APIs. Since high-quality instructions are expensive, the self instruction paradigm has gained popularity in recent days [20]. The main idea is to use different instruction templates to query GPT-3.5/4 and get more instructions from them automatically. Self instruction paradigm is designed based on in-context learning [21] ability of GPT-3.5/4. In this work, as shown in Fig. 3, we use some instances with the format <requirement, plan, script> as in-context examples to align GPT-3.5/4 and produce the instructions.

Instruction Collection. In alignment with the self-instruction paradigm, approximately 1500 instructions were devised for the purpose of instruction tuning. Given that GPT-3.5 and GPT-4 also generate erroneous data, some of the data were manually

created or adjusted. Concurrently, we conducted a thorough proofreading of the generated dataset to confirm the correct application of the EDA tools.

Instruction Fine-tuning. During the process of fine-tuning, each instruction comprises both a requirement and a response, with the response encompassing a plan and a script. To ensure an appropriate model sequence length, requirements and responses from the entire training set are concatenated, utilizing a unique token to demarcate these segments. An autoregressive objective is applied, zeroing out the loss on tokens derived from the user requirement, thus backpropagation is confined solely to response tokens.

Considering the imperative need for expedient training, as shown in Fig. 3, we adopt the QLoRA technique [15] to facilitate efficient fine-tuning. QLoRA leverages 4-bit Normalfloat (NF4) Quantization and Double Quantization techniques to achieve high-quality 4-bit fine-tuning. This approach is typically paired with paged optimizers to mitigate memory spikes during gradient checkpointing, thereby preventing potential Out-of-Memory errors. QLoRA’s efficacy allows us to guarantee the performance of instruction fine-tuning using LLMs on a large model scale (30B/70B), an accomplishment traditionally impeded by memory overhead constraints. QLoRA augments a linear projection in a transformer layer of LLMs through an additional factorized projection. As shown in the dash box of Fig. 3, given a linear projection, $\mathbf{Y} = \mathbf{X}\mathbf{W}$, the computation is as follows:

$$\mathbf{Y}^{\text{BF16}} = \mathbf{X}^{\text{BF16}} \text{doubleDequant}(c_1^{\text{FP32}}, c_2^{\text{k-bit}}, \mathbf{W}^{\text{NF4}}) + \mathbf{X}^{\text{BF16}} \mathbf{L}_1^{\text{BF16}} \mathbf{L}_2^{\text{BF16}}, \quad (1)$$

where \mathbf{W} is original weights and \mathbf{L}_1 and \mathbf{L}_2 are additional QLoRA weights. The final weights can be combined without extra inference costs. Here, the function $\text{doubleDequant}(c_1^{\text{FP32}}, c_2^{\text{k-bit}}, \mathbf{W}^{\text{k-bit}})$ can be obtained by:

$$\text{dequant}(\text{dequant}(c_1^{\text{FP32}}, c_2^{\text{k-bit}}), \mathbf{W}^{\text{4bit}}) = \mathbf{W}^{\text{BF16}}, \quad (2)$$

The matrix \mathbf{W} is of size NF4 and is quantized to 4-bit precision. The constant c_2 is of size FP8. We choose a blocksize of 64 for \mathbf{W} to achieve higher quantization precision, and a blocksize of 256 for c_2 to conserve memory. This choice is based on recent results showing that 8-bit quantization does not significantly degrade performance compared to 16-bit quantization.

IV. EXPERIMENTS

A. Setup

AutoMage is a fine-tuned model based on the Llama2-70B [8] architecture. For efficient fine-tuning on Llama2 using QLoRA, we implement a linear learning rate schedule using paged lion [22] 8-bit optimizer [23], initiating with a learning rate of 1×10^{-4} , no weight decay, a batch size of 128, and a sequence length of 4096 tokens. Ultimately, the model is fine-tuned for 15 epochs on $8 \times \text{A100}$ with 80G memory each.

During the inference phase, the user’s requirement prompts in natural language. These can be designed for a simple task (e.g. “Perform routing for the processor design on the asap7 platform.”) or delineate a broader, more general goal (e.g. “Please show me how to complete the design flow in the

script.”). Subsequently, API documentation is appended to the user’s prompt before being processed by AutoMage. The output of AutoMage is the executable script.

As for evaluation, we consider notable LLMs including Claude2 [6], GPT-3.5 [4], and GPT-4 [5] as our baselines for performance assessment. To ensure a comprehensive comparison, we utilize different LLMs as the core controllers for our autonomous agent, ChatEDA. The target API is a simplified python wrapper of OpenROAD [1].

B. Evaluation of LLMs

An objective evaluation system was developed to assess the task planning and script generation capabilities of LLMs designed for automated script generation. The system uses a three-tiered grading scheme, with Grade A representing the highest achievement. Grade A is awarded to LLMs that demonstrate coherent task planning and generate accurate scripts. Grade B indicates respectable but imperfect performance, assigned to LLMs that plan logically but falter in script generation. Grade C denotes failure in both planning and scripting.

During the evaluation process, we provide 50 cases including various tasks for comprehensive inspection. As summarized in Figure 4, our proposed AutoMage achieved the best performance, correctly earning Grade A for 88% of test cases. This significantly exceeds the 58% Grade A attained by the next highest performer, GPT-4. While GPT-4 exhibited reasonably strong capabilities, it struggled to differentiate between lower quality responses, assigning Grade B and C more evenly at 34% and 8% respectively. In contrast, AutoMage reliably identified the highest grade-worthy responses. The other LLMs, Claude2 and GPT-3.5, performed noticeably worse than AutoMage, with Claude2 achieving 44% Grade A and GPT-3.5 only 32% Grade A. This highlights their comparative deficiencies. In summary, AutoMage substantially outperformed GPT-4 and other LLMs by consistently assigning appropriate high grades. The considerable margin between AutoMage and second-place GPT-4 demonstrates the advances of our method in grading performance. Our experiment affirms AutoMage as the current state-of-the-art LLM for this evaluation task, which will be the most reliable controller of ChatEDA.

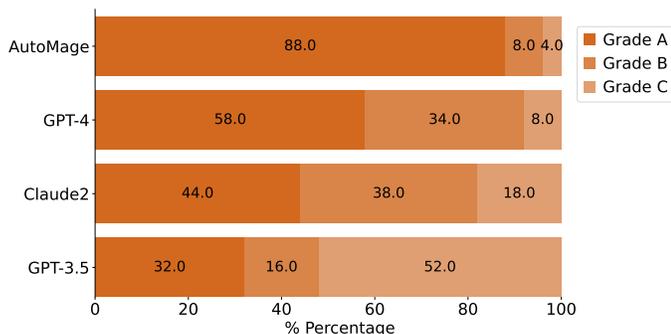


Fig. 4 Evaluation results for AutoMage compared to other notable LLMs. AutoMage outperforms other notable LLMs by a significant margin in task planning and script generation

C. Case Studies

In this section, we analyze five tasks involving user interactions with EDA tools through a conversational interface. It is important to note that this exploration is not exhaustive due to the limitations inherent in this qualitative study. For instance, controlling all possible variables, such as the vast array of responses the LLM can generate for a given prompt, is beyond our reach. As a result, our conclusions are based on a selected group of samples deemed representative. Nonetheless, we contend that the description of these exemplars will enhance our understanding of the robust performance of AutoMage, which enhances the reliability of AutoMage powered ChatEDA. Given that GPT-4 is the best LLM in baselines during the script evaluation, we focus exclusively on our AutoMage and GPT-4. For each task, we first outline the user’s requirements and subsequently present the scripts generated by both models.

Task 1: Performance Evaluation. The initial task necessitates the model’s successful execution of the comprehensive process, inclusive of the evaluation phase. This task underscores the fundamental application of the LLMs and their usage sequence in relation to the API interfaces. It is evident that both AutoMage and GPT-4 have completed this task satisfactorily. The output by GPT-4 is omitted since it is very similar to what AutoMage writes.

USER: I want to test the area and power performance of the design “leo” on “sky130” setting core utilization is 60%. I need to perform cts, routing, placement, and so on.
 AUTOMAGE

```

# Instantiate chateda
eda = chateda()
# Setup EDA environment
eda.setup(design_name="leo", platform="sky130")
# Run synthesis
eda.run_synthesis()
# Perform floorplanning with core utilization of 60%
eda.floorplan(core_utilization=60)
### INTERMEDIATE FLOW FUNCTION CALLS ARE OMITTED DUE TO PAGE LIMIT ###
# Generate the final report
eda.final_report()
# Evaluation
final_performance = eda.get_metric("final", ["area", "power"])

```

Task 2: Parameter Grid Search. The second task necessitates the execution of a grid search by LLMs, exhibiting a significant reliance on logic, including the traversal of parameters. This task further scrutinizes the LLM’s comprehension of each API argument. Evidently, both AutoMage and GPT-4 have performed superbly, demonstrating their logical prowess and exemplary understanding of API arguments.

USER: Your task is to grid search on or the design “how” on “gf180” platform for parameters core_utilization, clk period, and placement density.
 AUTOMAGE

```

# Define grid search parameters
core_utils = [60, 70, 80] # %
clk_periods = [2, 3, 4] # ns
densities = [0.6, 0.7, 0.8]
for core_util in core_utils:
    for clk_period in clk_periods:
        for density in densities:
            eda = chateda()
            # Setup
            eda.setup("how", "gf180", verilog="how.v")
            ### INTERMEDIATE FLOW FUNCTION CALLS ARE OMITTED DUE TO PAGE LIMIT ###
            # Finishing
            eda.final_report()

```

GPT-4

```
import numpy as np
eda = chateda.chateda()
# Setup design
eda.setup(design_name="how", platform="gf180")
# Parameter range
core_utilization_range = np.arange(0.1, 1.0, 0.1)
clock_period_range = np.arange(1, 11, 1)
placement_density_range = np.arange(0.1, 1.0, 0.1)
# Results
results = []
# Grid search over parameters
for core_utilization in core_utilization_range:
    for clock_period in clock_period_range:
        for placement_density in placement_density_range:
            eda.run_synthesis(clock_period=clock_period)
            eda.final_report()
            # Get metrics
            tns = eda.get_metric(stage="final", metrics=["tns"])
            # Store results
            results.append({
                "core_utilization": core_utilization,
                "clock_period": clock_period,
                "placement_density": placement_density,
                "tns": tns,
            })
### OTHER METRICS ARE OMITTED DUE TO PAGE LIMIT ###
```

Task 3: Parameter Tuning Method. The ensuing task necessitates the model to furnish a parameter-tuning solution. An investigation into the LLM’s application of the tune function simultaneously scrutinizes its capacity to offer an appropriate parameter space. Given the lack of adequate priors, LLMs can’t ascertain an exact parameter range. Nonetheless, it is observable that AutoMage, compared with GPT-4, leans more towards generating a complex search space, endowing AutoMage’s parameter tuning strategy with enhanced potential. TABLE I shows the result of parameter tuning. With the support of DSE engines, the design implemented by LLM-generated scripts outperforms default parameter settings for the platform.

USER: For the design “aes” on “nangate45” platform, fix clock period to be 5. Write me a script to optimize area and power using the parameter tuning method.

AUTOMAGE

```
def tuning_func(core_utilization, core_aspect_ratio, core_margins,
                macro_place_halo, macro_place_channel, density, tns_end_percent):
    eda = chateda()
    ### INTERMEDIATE FLOW FUNCTION CALLS ARE OMITTED DUE TO PAGE LIMIT ###
    eda.final_report()
    # Fetch and return metrics
    metrics = eda.get_metric("final", ["area", "power"])
    return metrics
param_space = {
    "core_utilization": {"minmax": [60, 90], "step": 5},
    "core_aspect_ratio": {"minmax": [0.8, 1.2], "step": 0.1},
    "core_margins": {"minmax": [8, 12], "step": 1},
    "macro_place_halo": {"minmax": [5, 9], "step": 1},
    "macro_place_channel": {"minmax": [7, 11], "step": 1},
    "density": {"minmax": [0.6, 0.9], "step": 0.05},
    "tns_end_percent": {"minmax": [30, 50], "step": 5}
}
# Execute tuning
tune(tuning_func, param_space)
```

GPT-4

```
eda = chateda()
eda.setup(design_name="aes", platform="nangate45")
# Define the function to run the flow and return area and power
def run_flow(param):
    # Synthesis
    eda.run_synthesis(clock_period=5)
    ### INTERMEDIATE FLOW FUNCTION CALLS ARE OMITTED DUE TO PAGE LIMIT ###
    eda.final_report()
    # Get area and power metrics
    area = eda.get_metric(stage="final", metrics=["area"])
    power = eda.get_metric(stage="final", metrics=["power"])
    # Return area and power as the result of the flow for tuning
    return area + power
# Define the parameter range for tuning
param_range = {
    "core_utilization": {"minmax": [0, 100], "step": 10},
    "density": {"minmax": [0, 1], "step": 0.1},
    "tns_end_percent": {"minmax": [0, 100], "step": 10},
}
# Perform tuning
tune(run_flow, param_range)
```

Task 4: Customized Optimization. Customized Optimization necessitates parameter tuning in the model, in accordance with the search space stipulated by the user. This process rigorously assesses the model’s comprehension of the tune function and each API interface parameter’s significance. It is evident that AutoMage aptly comprehends the requirement for user-customized parameter tuning, whereas GPT-4 demonstrates a deficient understanding of the concept of parameter combination optimization. This reveals a more profound understanding of the EDA tools’ interface and the logic of parameter tuning in AutoMage than in GPT-4.

USER: Please provide an optimal digital layout for our “high_end_gpu” project in the “nangate45” technology. Follow these steps:

- 1) Keep the clock period static at 5 during the synthesis.
- 2) At the floorplan stage, adjust only the core utilization, ranging it from 60% to 85% with a step of 5% and keep the rest of the parameters as their default values.
- 3) At the placement stage, try adjusting the density from 0.55 to 1 with an increment of 0.05.
- 4) At the CTS stage, fix from 30% to 60% of the violating paths with a step of 5%.

Lastly, gather metrics for chip area and power consumption once routing is finished.

AUTOMAGE

```
def tuning_func(core_utilization, density, tns_end_percent):
    eda = chateda()
    # Set up
    eda.setup("high_end_gpu", "nangate45")
    ### INTERMEDIATE FLOW FUNCTION CALLS ARE OMITTED DUE TO PAGE LIMIT ###
    # Perform routing
    eda.global_route()
    eda.detail_route()
    # Get metrics
    metrics = eda.get_metric("route", ["area", "power"])
    return metrics
# Define parameter space
params = {
    "core_utilization": {"minmax": [60, 85], "step": 5},
    "density": {"minmax": [0.55, 1], "step": 0.05},
    "tns_end_percent": {"minmax": [30, 60], "step": 5},
}
# Execute tuning
tune(tuning_func, params)
```

GPT-4

```
# Initialize chateda
eda = chateda()
# Step 1: Setup
eda.setup(design_name="high_end_gpu", platform="nangate45")
# Step 2: Synthesis
eda.run_synthesis(clock_period=5)
# Step 3: Floorplan with parameter tuning
def floorplan(core_utilization):
    eda.floorplan(core_utilization=core_utilization)
    performance = eda.get_metric(stage="floorplan", metrics=["area", "power"])
    return performance
params = {"core_utilization": {"minmax": [60, 85], "step": 5}}
tune(floorplan, params)
# Step 4: Placement with parameter tuning
def placement(density):
    eda.placement(density=density)
    performance = eda.get_metric(stage="place", metrics=["area", "power"])
    return performance
params = {"density": {"minmax": [0.55, 1], "step": 0.05}}
tune(placement, params)
# Step 5: CTS with parameter tuning
def cts(tns_end_percent):
    eda.cts(tns_end_percent=tns_end_percent)
    performance = eda.get_metric(stage="cts", metrics=["area", "power"])
    return performance
params = {"tns_end_percent": {"minmax": [30, 60], "step": 5}}
tune(cts, params)
### INTERMEDIATE FLOW FUNCTION CALLS ARE OMITTED DUE TO PAGE LIMIT ###
# Step 8: Final report and get metrics
eda.final_report()
chip_area = eda.get_metric(stage="final", metrics=["area"])
power_consumption = eda.get_metric(stage="final", metrics=["power"])
print("Chip area: ", chip_area)
print("Power consumption: ", power_consumption)
```

Design	clk	Default			GPT3.5			GPT-4			AutoMage			
		WNS	Power	Area	WNS	Power	Area	WNS	Power	Area	WNS	Power	Area	
aes	0.82	-0.149	233.6	42059.4	-0.115	232.0	42059.4	-0.167	236.3	52828.7	-0.150	250.7	34977.7	
gcd	0.74	-0.555	1.723	1355.54	> 0	1.230	983.14	> 0	1.230	983.14	> 0	1.231	746.13	
ibex	2.8	-0.082	145.2	56237.5	-0.095	139.0	56237.5	-0.095	139.0	56237.5	-0.095	139.0	56237.5	
jpeg	1.7	-0.058	636.6	191214	-0.030	628.6	191214	-0.030	628.6	191214	-0.030	628.6	191214	
Average Power*Area (Scaled)			1.31			1.10			1.17			1.00		

TABLE I Results of parameter tuning for power and area optimization (Task 3). The script generated by GPT-3.5 contains several semantic errors, which are manually fixed before running experiments.

Task 5: Clock Period Minimization. Throughout the process of instruction tuning, we guided AutoMage on utilizing the API interface based on our specific needs. However, we did not teach it to modify the parameters using self-feedback derived from the evaluation outcomes. In the final test case, we scrutinized how the LLM adjusts the API arguments via self-feedback to achieve the user-requested evaluation results. Intriguingly, AutoMage successfully accomplished the provided test case, an endeavor in which GPT-4 failed. This underscores AutoMage’s profound comprehension of utilizing EDA tools.

USER: Try to find out the smallest valid clock period for the design “leon” on “asap7” platform. Note that a clock period is valid only if the “wns” metric at the final stage is non negative.

AUTOMAGE

```
def find_smallest_valid_clock_period(clock_period):
    eda_tool = chateda()
    eda_tool.setup(design_name="leon", platform="asap7")
    ### INTERMEDIATE FLOW FUNCTION CALLS ARE OMITTED DUE TO PAGE LIMIT ###
    eda_tool.final_report()
    # Get the wire congestion at the final stage
    final_metrics = eda_tool.get_metric("final", ["wns"])
    if final_metrics >= 0:
        return True
    else:
        return False
# iterate over different clock periods and check which one gives a valid result
clock_periods = [1, 2, 3, 4, 5] # replace with desired clock periods
smallest_valid_clock_period = 0
for clock_period in clock_periods:
    if find_smallest_valid_clock_period(clock_period):
        smallest_valid_clock_period = clock_period
        break
print(smallest_valid_clock_period)
```

GPT-4

```
# create an instance of chateda
eda = chateda.chateda()
# define the function to be used for tuning
def run_flow(clock_period):
    eda.setup(design_name="leon", platform="asap7")
    ### INTERMEDIATE FLOW FUNCTION CALLS ARE OMITTED DUE TO PAGE LIMIT ###
    eda.final_report()
    wns = eda.get_metric(stage="final", metrics=["wns"])
    return wns
# define the parameter space for tuning
param = {"clock_period": {"minmax": [1, 10], "step": 0.1}}
# run tuning
tune(func=run_flow, param=param)
```

V. CONCLUSION

Interfacing EDA tools is essential for unleashing circuit design productivity. In this work, we propose to an LLM-powered autonomous agent for EDA, which enables a conversational interface for designers to interact with the design flow. Technically, ChatEDA integrates a fine-tuned AutoMage, which orchestrates the design flow through task planning, script generation, and task execution. ChatEDA handles various user requirements well, outperforming other LLM models like GPT-4 and so on. We hope this work could inspire next-generation EDA tool evolution.

REFERENCES

- [1] T. Ajayi and D. Blaauw, “Openroad: Toward a self-driving, open-source digital layout implementation tool chain,” in *Proceedings of Government Microcircuit Applications and Critical Technology Conference*, 2019.
- [2] X. Li *et al.*, “ieda: An open-source intelligent physical implementation toolkit and library,” in *International Symposium of Electronics Design Automation (ISED)*, 2023.
- [3] P. Chen, D. A. Kirkpatrick, and K. Keutzer, “Scripting for eda tools: a case study,” in *Proceedings of the IEEE 2001. 2nd International Symposium on Quality Electronic Design*. IEEE, 2001, pp. 87–93.
- [4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” in *Proc. NIPS*, 2020.
- [5] OpenAI, “Gpt-4 technical report,” 2023.
- [6] A. Team, “Claude2,” <https://www.anthropic.com/index/claude-2>, 2023.
- [7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint*, 2023.
- [8] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint*, 2023.
- [9] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, “Emergent abilities of large language models,” *Journal of Machine Learning Research*, 2022.
- [10] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, “Self-instruct: Aligning language model with self generated instructions,” *arXiv preprint*, 2022.
- [11] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan, “Chatlaw: Open-source legal large language model with integrated external knowledge bases,” *arXiv preprint*, 2023.
- [12] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, “Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge,” *arXiv preprint*, 2023.
- [13] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, “Finetuned language models are zero-shot learners,” in *Proc. ICLR*, 2022.
- [14] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, “Judging llm-as-a-judge with mt-bench and chatbot arena,” *arXiv preprint*, 2023.
- [15] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” *arXiv preprint*, 2023.
- [16] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah, “Orca: Progressive learning from complex explanation traces of gpt-4,” *arXiv preprint*, 2023.
- [17] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, “Toolformer: Language models can teach themselves to use tools,” *arXiv preprint*, 2023.
- [18] A.-G. Team, “Auto-gpt: An autonomous gpt-4 experiment,” <https://github.com/Significant-Gravitas/Auto-GPT>, 2023.
- [19] B. Team, “Babyagi,” <https://github.com/yoheinakajima/babyagi>, 2023.
- [20] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” in *Proc. NIPS*, 2022.
- [21] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, “A survey for in-context learning,” *arXiv preprint*, 2022.
- [22] X. Chen, C. Liang, D. Huang, E. Real, K. Wang, Y. Liu, H. Pham, X. Dong, T. Luong, C.-J. Hsieh *et al.*, “Symbolic discovery of optimization algorithms,” *arXiv preprint*, 2023.
- [23] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer, “8-bit optimizers via block-wise quantization,” *arXiv preprint*, 2021.