

Chip2System

LLM-gestütztes Co-Design

Julian Withöft

julian.withoeft@tu-dortmund.de

Werner John

Philip Vargehese Modayil

Emre Ecik

Nima Ghafarian Shoaee

Bekhzodkhon Khamdamov

Domenik Helms

Ralf Brüning

Jürgen Götze

tu technische universität
dortmund

DT/PL Arbeitsgebiet Datentechnik
Information Processing Lab



PYRAMIDE 2525
Elektrische Entwicklung
KI-Anwendungen



ZUKEN[®]

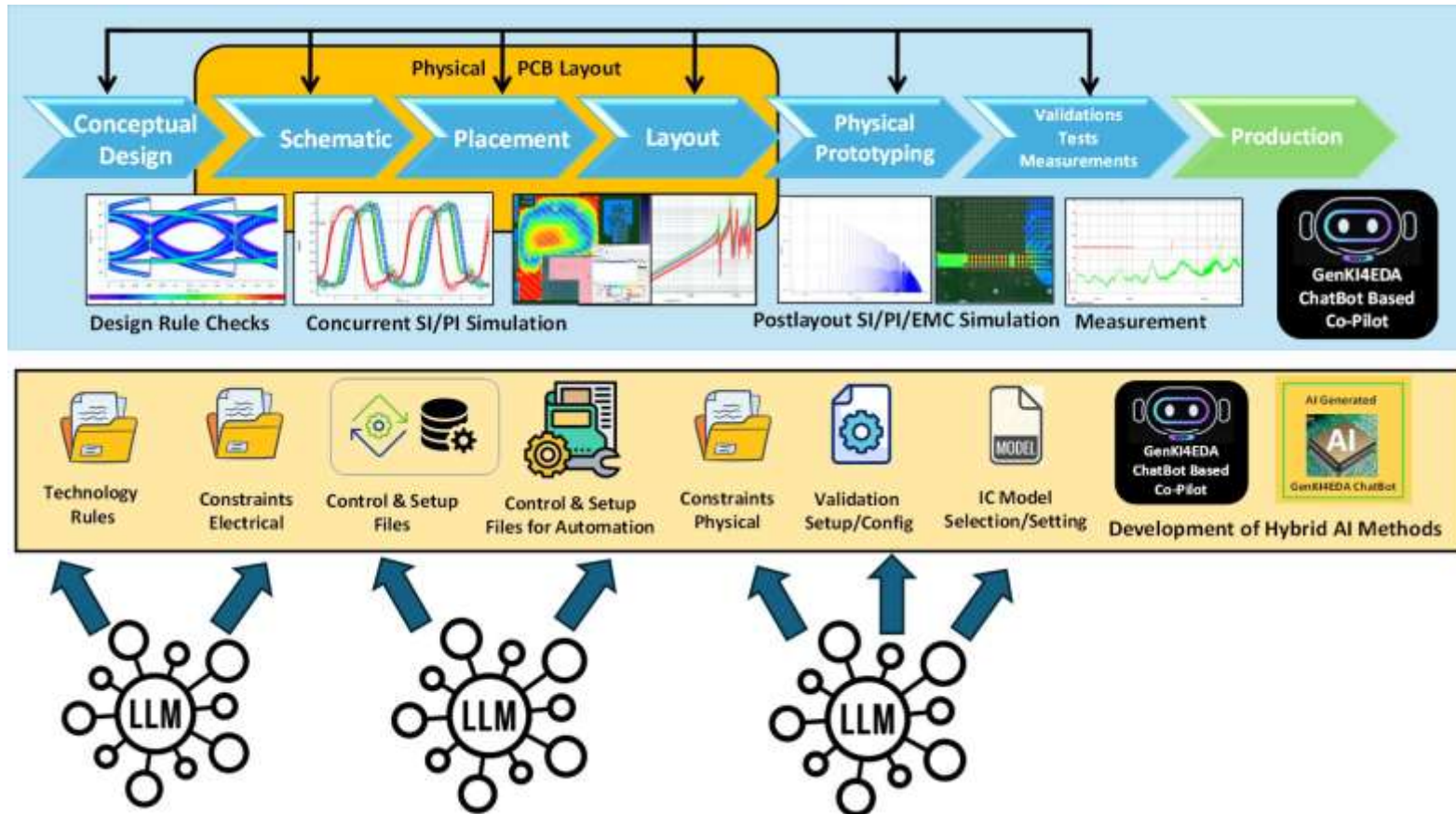
Das diesem Beitrag zugrundeliegende Vorhaben KI4BoardNet wird vom Bundesministerium für Bildung und Forschung (BMBF) (Förderprogramm Mannheim) unter dem Kennzeichen (Förderkennzeichen: 16ME0779/TUDO - 16ME0777/IDMT) gefördert. Für den Inhalt dieser Publikation sind die Autoren verantwortlich.

Gliederung

- Motivation
- Rasa Intent-based Bot
- LLM als Fallback
- Fine-Tuning eines Domänen-spezifischen LLMs
- LLM Agents
- Anwendungsbeispiele
- Zusammenfassung und Ausblick

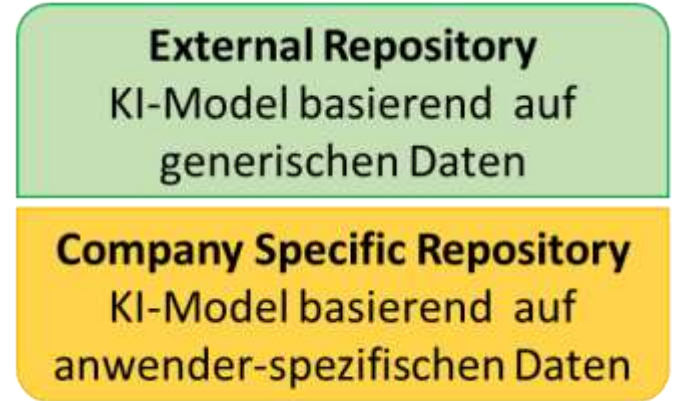
Co-Pilot Motivation für Chip2System Entwurfsaufgaben

- Chatbot-gestützte Design-Assistenz (Co-Piloting) in jedem Entwurfschritt mittels Nutzung LLM-generierter Control Sets, welche Regeln/Constraints für individuelle Design Sets enthalten
- Chatbot liefert Design-Empfehlungen und auch Input für Entwicklungsdaten basierend auf Design-Kontext



Co-Pilot Motivation für Chip2System Entwurfsaufgaben

- Sicherstellung eines umfassenden IP-Schutzes →
- Ziel ist Entwicklung hybrider KI-Methoden: Wissen (Regeln) mit Lerndaten (z.B. LLM) kombinieren → Unterbindung von Halluzinationen.
- Implementation multimodaler LLMs
 - Training mit verschiedenen Datenquellen (Text; Bild/Video; Sprache).
 - Analyse von Schaltungsentwürfen direkt aus Bild- oder Simulationsdaten.
- Chatbot-Wissensbasis: Strukturiertes Systementwurfswissen (Vektor Datenbank + MetaDaten)
 - Chatbot Dokumentenanalyse/Verarbeitung für die EDA-Wissensdomänen IC + PCB (EMC/SI/PI) + Bordnetz.
 - Chatbot-Infrastruktur (EDA-Wissensdomänen IC + PCB (EMC/SI/PI) + Bordnetz + Konzeptions- und Entwicklungsprozesse).
- Integration von Simulations- und AI-APIs (Agenten)
 - EDA-Werkzeuge (Simulatoren + Analyse).
 - KI-Modelle.
 - LLM-Modelle.



Rasa Intent-Based Bot

- Fokus auf sehr kleine Wissensdomäne: Decoupling Capacitors
- Natural Language Understanding (NLU)
- Antwort kommt direkt aus einer Datenbank



Einfach, Strukturiert, Effizient, Kontrolliert



Wenig Flexibel, Nicht Skalierbar, Limitiert auf Datenbank

Workflow of Intent Based Model



Quelle: <https://rasa.com>



LLM Fallback

- Wenn Rasa-Confidence < Threshold → LLM (Meta-Llama-3-8B) wird für die Response verwendet



Jede Frage kann beantwortet werden, Externe Datenquellen können via RAG eingebunden werden, LLM läuft Lokal auf dem Server



Komplexität, Halluzinationen, LLM ist kein PCB/IC-Experte

tu SI/PI-Compliant PCB Design Assistant - Version 2: Rasa Intent-Based Bot + LLM (LLaMA 3) Fallback

User: Hi Chatbot

HAL: Hello! I'm a PCB design assistant. How can I help you today?

User: What are the different types of decoupling capacitors?

HAL: The different types are:

1. Electrolytic Capacitors
2. Aluminum Electrolytic Capacitors
3. Tantalum Capacitors
4. Surface Mount Ceramic Capacitors
5. Surface Mount MLCC (MultiLayer Ceramic Capacitors)
6. Film Capacitors

User: what are the different types of transmission lines on a pcb?

HAL: There are several types of transmission lines on a PCB (Printed Circuit Board), each with its own characteristics and applications. Here are some of the most common types:

1. Microstrip transmission line: A microstrip transmission line consists of a thin layer of conductive material (e.g., copper) on one side of a dielectric material (e.g., FR4) and is used for high-frequency applications.
2. Stripline transmission line: A stripline transmission line consists of a conductive material (e.g., copper) sandwiched between two dielectric materials (e.g., FR4) and is used for high-frequency applications.
3. Coplanar waveguide (CPW) transmission line: A CPW transmission line consists of a conductive material (e.g., copper) on one side of a dielectric material (e.g., FR4) and is used for high-frequency applications.
4. Edge-coupled transmission line: An edge-coupled transmission line consists of two conductive materials (e.g., copper) on opposite sides of a dielectric material (e.g., FR4) and is used for low-frequency applications.
5. Shielded transmission line: A shielded transmission line consists of a conductive material (e.g., copper) wrapped around a dielectric material (e.g., FR4) and is used for high-frequency applications.
6. Twin-lead transmission line: A twin-lead transmission line consists of two parallel conductive materials (e.g., copper) separated by a dielectric material (e.g., FR4) and is used for low-frequency applications.
7. Coaxial transmission line: A coaxial transmission line consists of a conductive material (e.g., copper) wrapped around a central dielectric material (e.g., FR4) and is used for high-frequency applications.

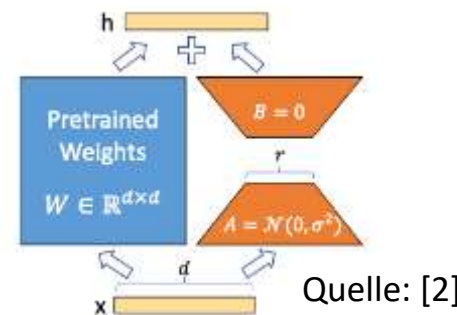
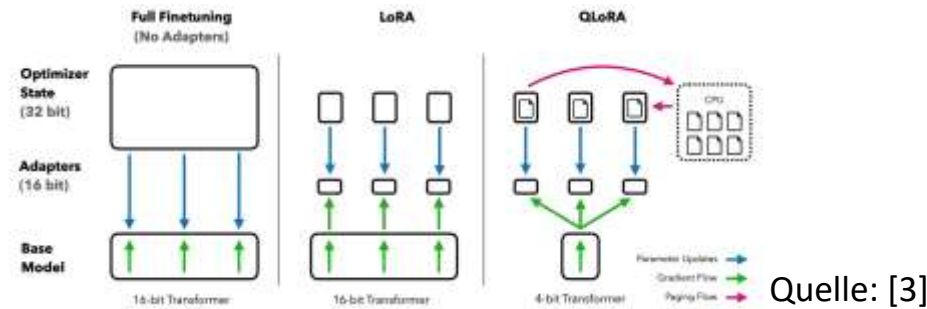
HAL: These are just a few examples of the different types of transmission lines that can be found on a PCB. The specific type of transmission line used will depend on the application and the frequency range of the signal being transmitted.

LLM Fine-Tuning

- Supervised Fine-Tuning → Spezialisierung des LLMs in einer Domäne
- LLaMA-Factory [1] vereinfacht den Fine-Tuning Prozess
- QLoRA wird verwendet (4-Bit Quantisierung) [2,3]



Quelle: <https://github.com/hiyoga/LLaMA-Factory>



- In [4] wurde gezeigt, dass bereits mit 1000 Samples eine gute Performance zu erreichen ist
- Unser Datensatz mit insgesamt 2247 Q&A manueller Samples aus der SI/PI/EMC Domäne → 2197 Training Samples, 50 Test Samples

[1] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma, "LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models," arXiv preprint arXiv:2403.13372, 2024.

[2] E.J.Hu, Y.Shen, P.Wallis, Z.Allen-Zhu, Y.Li, S.Wang, L.Wang, and W.Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.

[3] T. Detmners, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, 2023.

[4] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy, "LIMA: Less Is More for Alignment," arXiv preprint arXiv: 2305.11206, 2023.

Evaluation der Fine-Tuned LLMs

- LLM-as-a-Judge [5] Prinzip wird verwendet, um die Antworten der LLMs auf Basis des Testdatensatzes zu bewerten:

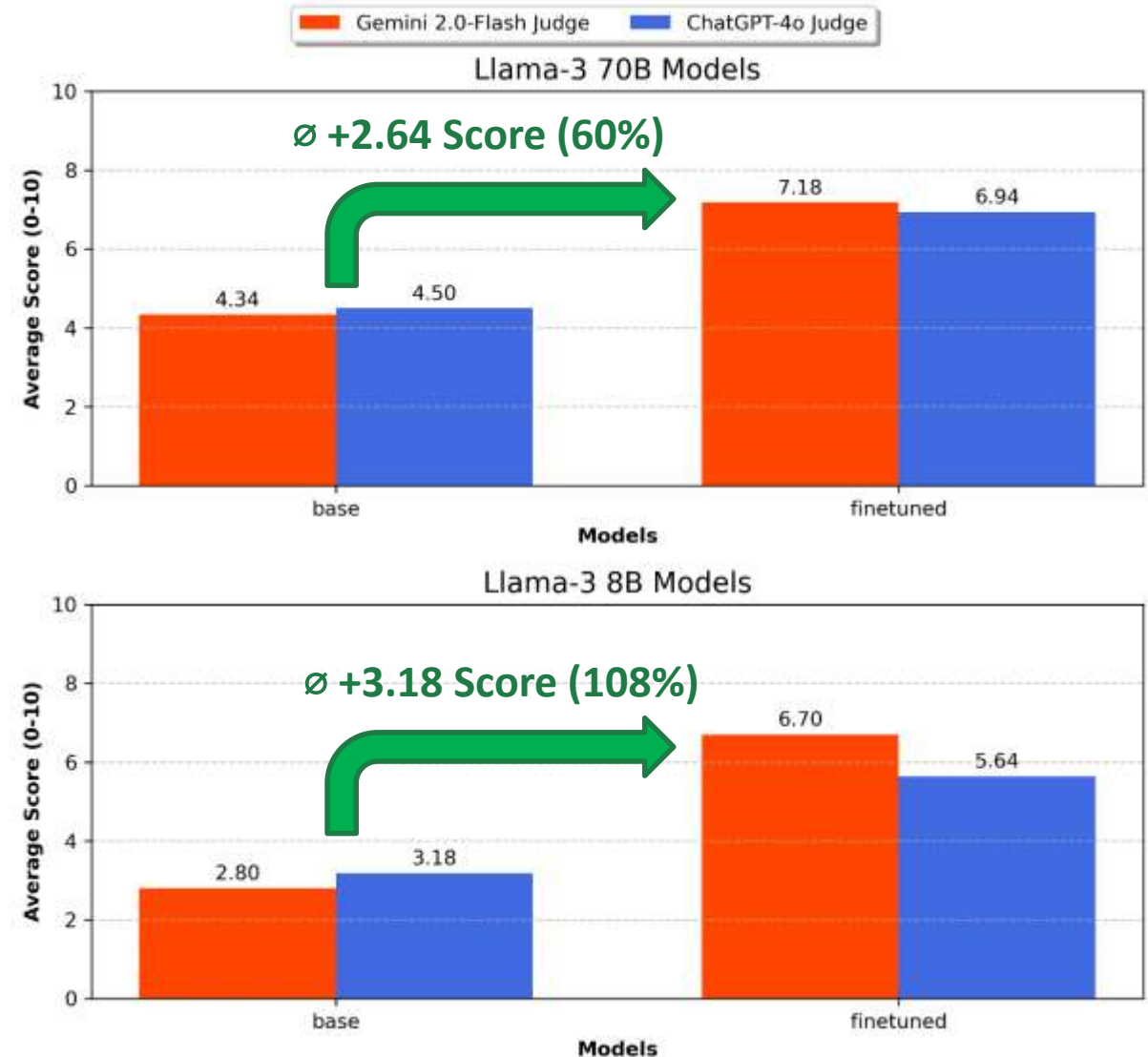
```
[Instruction]
Please act as an impartial judge and evaluate the quality of the
response provided by an AI assistant to the user question
displayed below. Your evaluation should consider correctness
and helpfulness. You will be given a reference answer and the
assistant's answer. Begin your evaluation by comparing the
assistant's answer with the reference answer. Identify and
correct any mistakes. Be as objective as possible. After providing
your explanation, you must rate the response on a scale of 1 to 10 by
strictly following format: "[[rating]]", for example: "Rating: [[5]]".

[Question]
{question}

[The Start of Reference Answer]
{ref_answer_1}
[The End of Reference Answer]

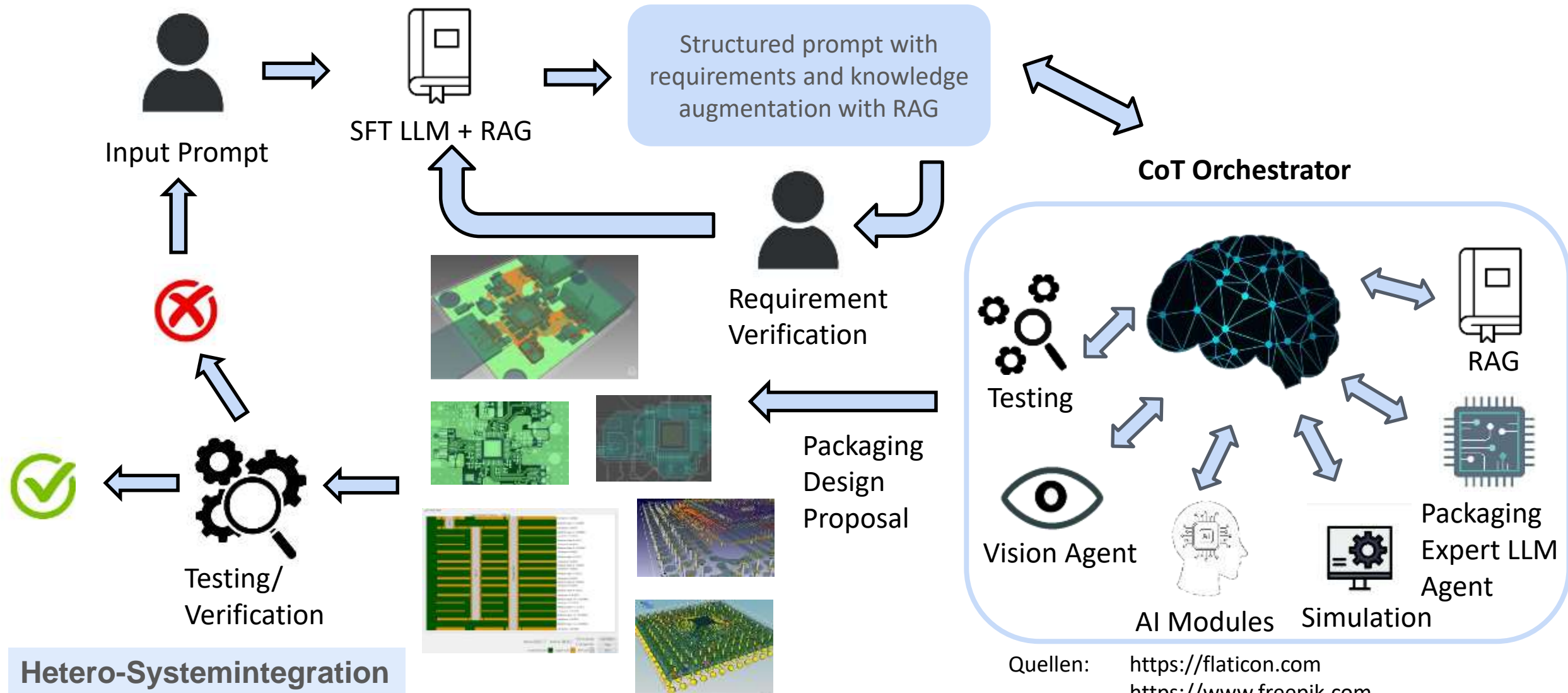
[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]
```

- Fine-Tuning führt eindeutig zu **besseren** Antworten der LLMs



[5] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," arXiv preprint arXiv:2306.05685, 2023.

LLM Agent Framework



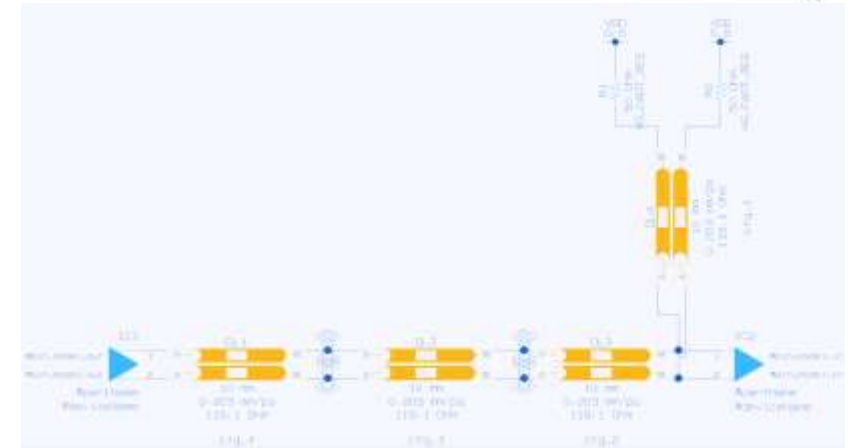
Quellen: <https://flaticon.com>
<https://www.freepik.com>

LLM Agent – Exemplarisches Anwendungsszenario

Stackup-Recommendation for Differential Pair Point-to-Point (P2P DDR4-Interconnect)

Agentic LLM Workflow (Exemplarisch):

- **User Input:** Datasheet TI ; Screenshot Schematic + **User Prompt:** I am designing a PCB and I am planning to use the AM64x IC (see attached datasheet). I have to route the DQS DDR4 signals as a differential pair point-to-point interconnect (see screenshot). Can you recommend a layer stackup for my design and on which layer to route my differential DQS signals specifically? Also what is the maximum transmission line length that is allowed and where should the vias be placed? Finally, what is the max tolerance of the transmission line parameters (i.e. width, spacing, etc.)?
- **SFT LLM + RAG** scans datasheet reformulates the requirements & RAG scans datasheet for additional information (i.e. Skew thresholds (here 0.4 ps), Eye mask, etc.)
- **CoT Orchestrator <Think>** : I have a differential DDR4 point-to-point interconnect. I need to specify layer stackup and answer design-specific questions. First I need more information regarding DDR4 → **Invoke Tool: RAG Agent**
- **RAG Agent:** I need further information on DDR4 so I will search for further information... Found JEDEC-DDR4 specification on local server → Scan and add to vector database.
- **<Think>** : I need to evaluate the schematic that the user provided to check what is currently implemented → **Invoke Vision Agent**
- **MM-LLM Vision Agent:** Analyzing Image.. Extracting Parameter Values.. Validating that it is a Differential Point-to-Point Topology.. Finished.. → Give back information to orchestrator
- **<Think>** : I need to optimize the vias placement, transmission line lengths and tolerances regarding transmission line parameters.. Search for AI Modules for this design task.. Found AI module for specific interconnect.. → **Invoke AI Module Agent**
- **AI Module Agent:** Searching for matching AI modules.. Found AI prediction module based on ANN model.. Running GA optimization.. Optimization finished..
- **<Think>** : I need to verify the design proposal of the AI module with some simulations. → **Invoke Simulation Agent**
- **Simulation Agent:** Generate *.sdf file.. Starting simulations via Zuken eCADSTAR.. Evaluating simulations.. Violations found – need to adjust vias and tolerances slightly.. Make adjustments to design proposal.. Finalize design proposal..
- **<Think>** : I have found a design proposal. Do I need to check anything else? If yes, go back into loop, if no, give design proposal back to user



Application Note AM64x/AM243x DDR Board Design and Layout Guidelines

TEXAS INSTRUMENTS

ABSTRACT
The goal of this document is to describe how to make the AM64x/AM243x DDR system implementation straightforward for all designers. The requirements have been distilled down to a set of layout and routing rules that allow designers to successfully implement a robust design for the topologies TI supports.

Table of Contents	
1 Overview	2
1.1 Board Designs Supported	2
1.2 General Board Layout Guidelines	2
1.3 PCB Stack-Up	3
1.4 Eye-mask Calculation	4
1.5 Velocity Compensation	5
2 DDR4 Board Design and Layout Guidance	6
2.1 DDR4 Introduction	6
2.2 DDR4 Device Implementations Supported	6
2.3 DDR4 Interface Schematics	7
2.4 Compatible JEDEC DDR4 Devices	16
2.5 Placement	13
2.6 DDR4 Keepout Region	11
2.7 VTT	11
2.8 Net Classes	11
2.9 DDR4 Signal Termination	12
2.10 VREF Routing	12
2.11 VTT	12
2.12 POD Interconnect	12
2.13 CK and ADDR_CTRL Topologies and Routing Guidelines	13
2.14 Data Group Topologies and Routing Guidelines	16
2.15 CK and ADDR_CTRL Routing Specification	17
2.16 Data Group Routing Specification	18
2.17 84 Swapping	20
3 LPDDR4 Board Design and Layout Guidance	21
3.1 LPDDR4 Introduction	21
3.2 LPDDR4 Device Implementations Supported	21
3.3 LPDDR4 Interface Schematics	22
3.4 Compatible JEDEC LPDDR4 Devices	23
3.5 Placement	23
3.6 LPDDR4 Keepout Region	24
3.7 Net Classes	24
3.8 LPDDR4 Signal Termination	24
3.9 LPDDR4 VREF Routing	25
3.10 LPDDR4 VTT	25
3.11 CK and ADDR_CTRL Topologies	25
3.12 Data Group Topologies	25
3.13 CK and ADDR_CTRL Routing Specification	27
3.14 Data Group Routing Specification	28
3.15 Channel, Byte, and 84 Swapping	28
4 Revision History	28

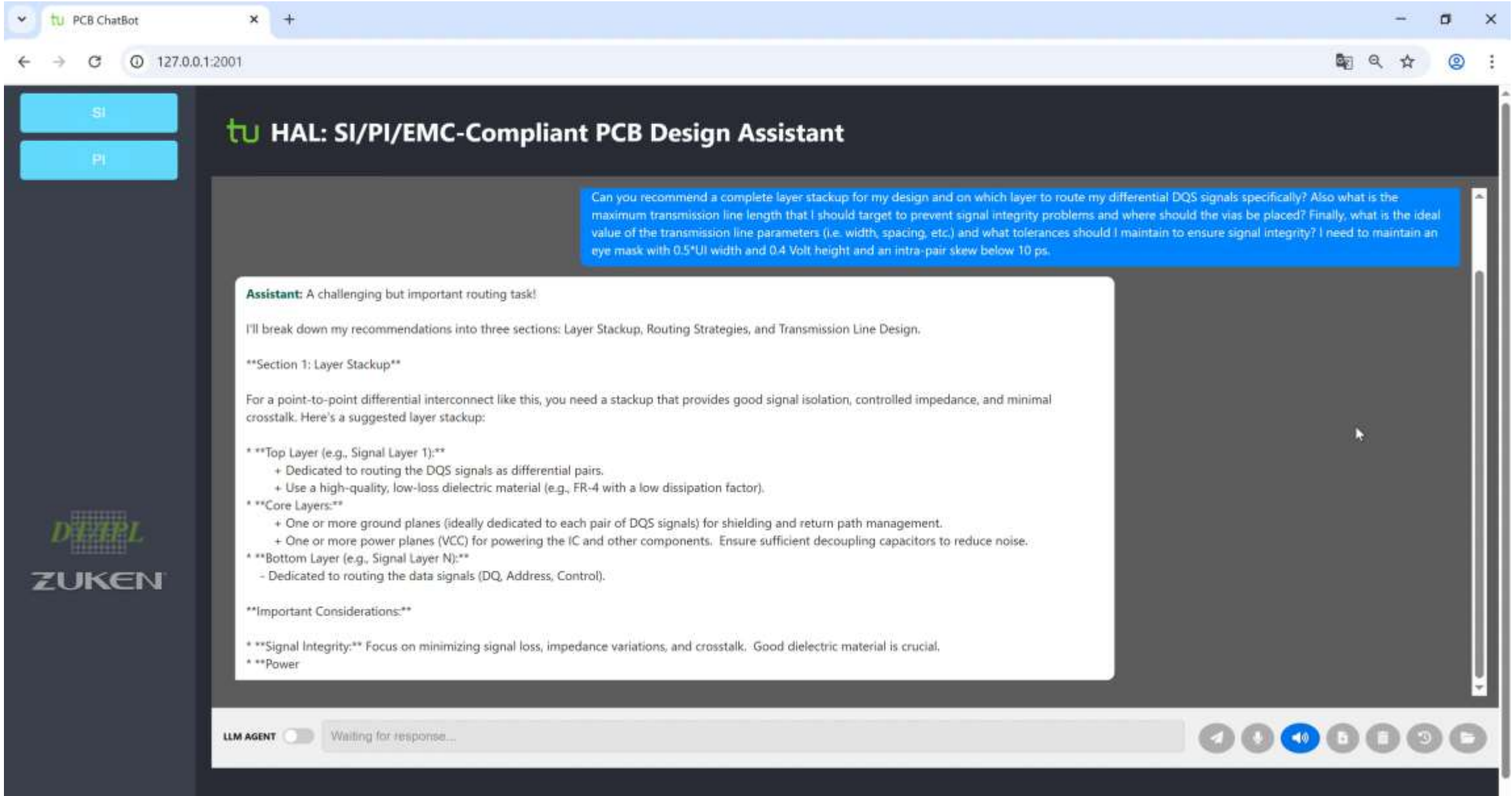
Anwendungsbeispiele #1 – Rasa SI Design Support

The screenshot shows a web browser with multiple tabs for 'PCB ChatBot'. The address bar shows '127.0.0.1:6015'. The interface is dark-themed with blue buttons. On the left, there are buttons for 'SI' and 'PI'. The main content area is titled 'Point-to-Point Topology' and includes several sections:

- Navigation:** Point-to-Point Topology (selected), Star Topology, Daisy Chain Topology.
- PCB Parameters:** Resistor RT: 57.96, Length L1: 18.81, Length L2: 1, Length L3: 18.81, Impedance L1: 60.03, Impedance L2: 77.16, Impedance L3: 64.34, Dielectric Constant: 4.
- IC Parameters:** Clock Frequency: 800, VCC Voltage: 1.5, Rise & Fall Time: 300, Jitter: 3, Resistor TX: 31, Die Capacitance TX: 5, Die Capacitance RX: 1.
- Simulation Tool:** SPICE (selected), eCADSTAR.
- Diagram:** A schematic diagram showing a signal path from V_{IC1} through a transmission line $Z_{0,1}$ (length l_1) to a junction. At the junction, there is a resistor R_T connected to $VCC/2$ and a transmission line $Z_{0,2}$ (length l_2) leading to V_{IC2} . A third transmission line $Z_{0,3}$ (length l_3) is also connected to the junction.
- AI Optimization Results:** A table with the following data:

Parameter	Value
RT	57.96 Ω
L1	18.81 mm
L2	1 mm
L3	18.81 mm
Z1	60.03 Ω
Z2	77.16 Ω
Z3	64.34 Ω

Buttons: 'Back to AI Optimization', 'Apply AI Results'.
- Eye Simulation:** A plot of Voltage V_{IC2} [V] vs Time [ns]. The plot shows two eye diagrams. Text above the plot: 'Eye Height: 938 mV : Eye Width: 410 ps'. Buttons: 'Transient Simulation', 'Eye Simulation'.



tu PCB ChatBot

127.0.0.1:2001

SI

PI

tu HAL: SI/PI/EMC-Compliant PCB Design Assistant

Can you recommend a complete layer stackup for my design and on which layer to route my differential DQS signals specifically? Also what is the maximum transmission line length that I should target to prevent signal integrity problems and where should the vias be placed? Finally, what is the ideal value of the transmission line parameters (i.e. width, spacing, etc.) and what tolerances should I maintain to ensure signal integrity? I need to maintain an eye mask with 0.5*UI width and 0.4 Volt height and an intra-pair skew below 10 ps.

Assistant: A challenging but important routing task!

I'll break down my recommendations into three sections: Layer Stackup, Routing Strategies, and Transmission Line Design.

****Section 1: Layer Stackup****

For a point-to-point differential interconnect like this, you need a stackup that provides good signal isolation, controlled impedance, and minimal crosstalk. Here's a suggested layer stackup:

- **Top Layer (e.g., Signal Layer 1):****
 - + Dedicated to routing the DQS signals as differential pairs.
 - + Use a high-quality, low-loss dielectric material (e.g., FR-4 with a low dissipation factor).
- **Core Layers:****
 - + One or more ground planes (ideally dedicated to each pair of DQS signals) for shielding and return path management.
 - + One or more power planes (VCC) for powering the IC and other components. Ensure sufficient decoupling capacitors to reduce noise.
- **Bottom Layer (e.g., Signal Layer N):****
 - Dedicated to routing the data signals (DQ, Address, Control).

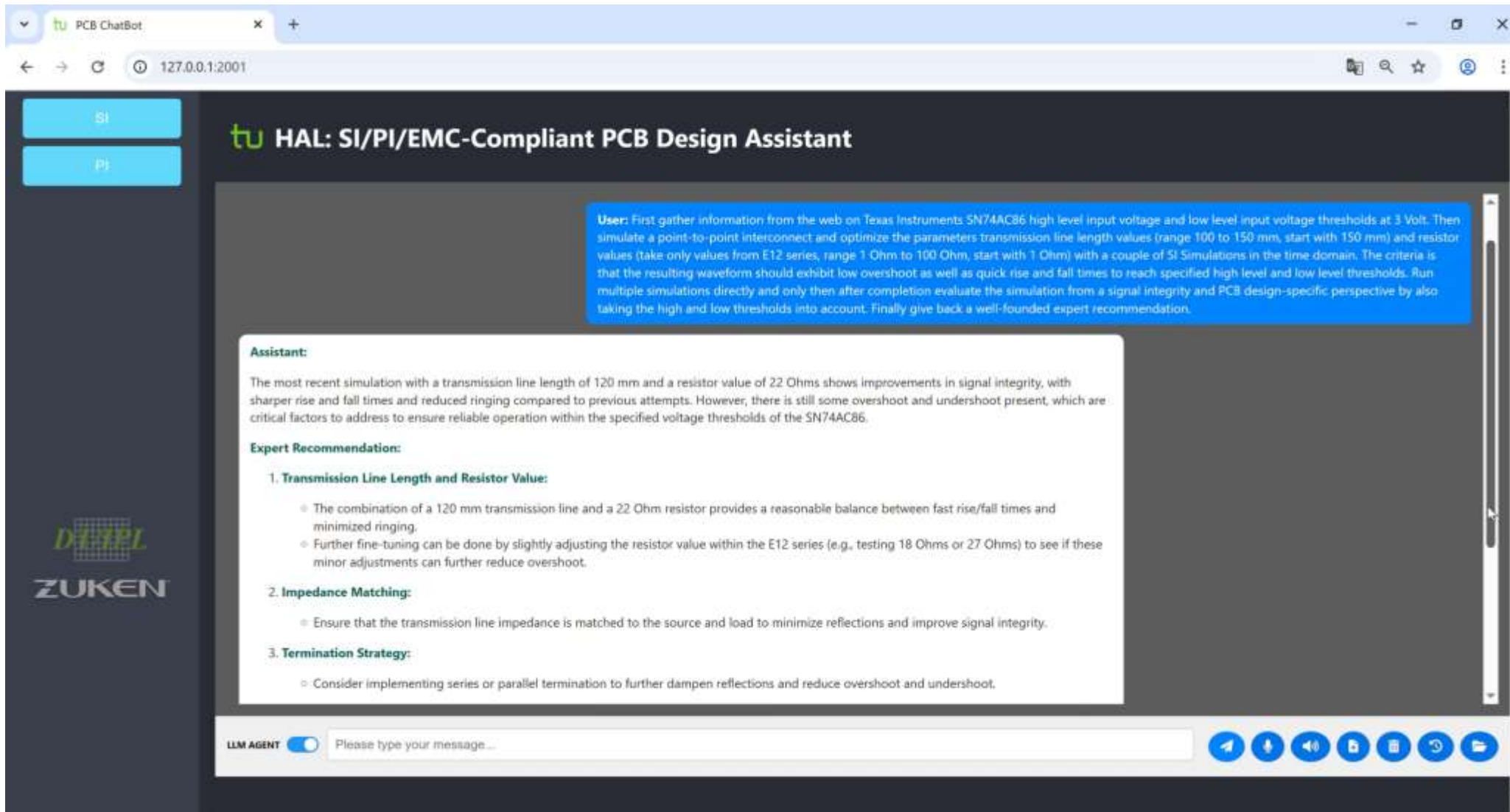
****Important Considerations:****

- **Signal Integrity:**** Focus on minimizing signal loss, impedance variations, and crosstalk. - Good dielectric material is crucial.
- **Power**

LLM AGENT Waiting for response...

Navigation icons: back, forward, search, star, share, refresh, close

Anwendungsbeispiel #3 – LLM Agent Task



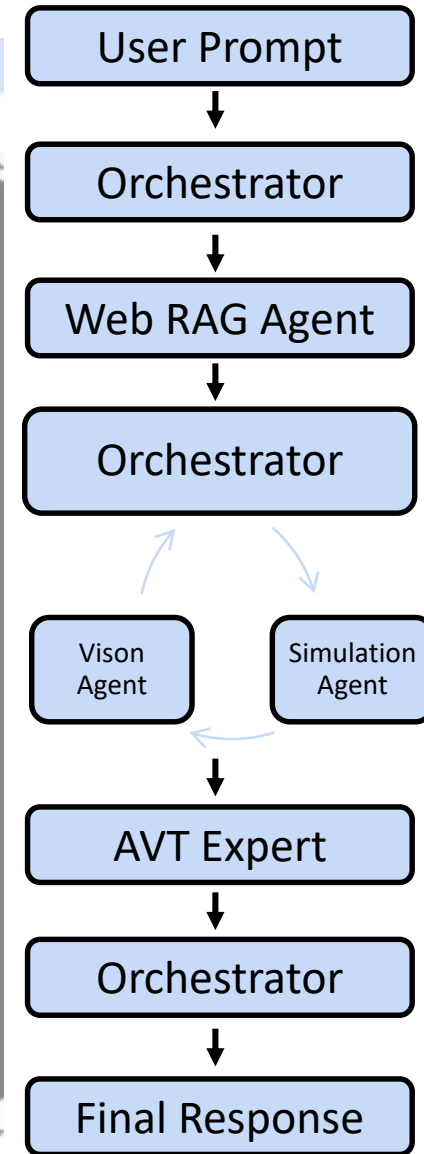
User: First gather information from the web on Texas Instruments SN74AC86 high level input voltage and low level input voltage thresholds at 3 Volt. Then simulate a point-to-point interconnect and optimize the parameters transmission line length values (range 100 to 150 mm, start with 150 mm) and resistor values (take only values from E12 series, range 1 Ohm to 100 Ohm, start with 1 Ohm) with a couple of SI Simulations in the time domain. The criteria is that the resulting waveform should exhibit low overshoot as well as quick rise and fall times to reach specified high level and low level thresholds. Run multiple simulations directly and only then after completion evaluate the simulation from a signal integrity and PCB design-specific perspective by also taking the high and low thresholds into account. Finally give back a well-founded expert recommendation.

Assistant:

The most recent simulation with a transmission line length of 120 mm and a resistor value of 22 Ohms shows improvements in signal integrity, with sharper rise and fall times and reduced ringing compared to previous attempts. However, there is still some overshoot and undershoot present, which are critical factors to address to ensure reliable operation within the specified voltage thresholds of the SN74AC86.

Expert Recommendation:

- Transmission Line Length and Resistor Value:**
 - The combination of a 120 mm transmission line and a 22 Ohm resistor provides a reasonable balance between fast rise/fall times and minimized ringing.
 - Further fine-tuning can be done by slightly adjusting the resistor value within the E12 series (e.g., testing 18 Ohms or 27 Ohms) to see if these minor adjustments can further reduce overshoot.
- Impedance Matching:**
 - Ensure that the transmission line impedance is matched to the source and load to minimize reflections and improve signal integrity.
- Termination Strategy:**
 - Consider implementing series or parallel termination to further dampen reflections and reduce overshoot and undershoot.



Zusammenfassung und Ausblick

- Regelbasierte Aufgaben mittels KI-Verfahren automatisieren (LLM/Generative KI) → Entscheidender Vorteil im Vergleich zu **konventionellen EDA-Methoden** (Regelüberprüfung mittels Simulation und geometrischen Verfahren)
- Dieser Ansatz trägt dazu bei, **manuell-gesteuerte Entwurfsschritte** zu **minimieren** und so zu erreichen, dass die Entwickler (Ingenieure) sich auf Innovationen und damit komplexe Problemlösungen konzentrieren können.
- **LLM-gestützte Co-Pilots** besitzen das Potenzial **EDA Prozesse effizienter** zu **gestalten** und den **Entwickler** bei ausgewählten **Entwurfsaufgaben** zu **unterstützen** → AVT/Packaging-Trends: PCB - Advanced Packaging - Chiplet - IC-Substrate – EDA AddOns
- **Fine-Tuning** ermöglicht die **Spezialisierung** von LLMs in einer **spezifischen Domäne** → Hier: **SI/PI/EMC** im **Leiterplattendesign/AVT**
- **Zukünftig** werden **insbesondere LLM Agents** und deren **Integration** in **EDA/CAD-Tools** relevant sein, um bestimmte Aufgaben mit **Tools** (Simulation, Optimierung, Requirements, RAG, KI-Module ...) innerhalb eines Loops durchzuführen und den Entwickler so noch besser unterstützen zu können.
- Forschungsansatz: *Hybride KI-Methoden* ↔ **Wissen (Regeln)** mit **Lerndaten** (z.B. LLM) kombinieren (Entwicklung *ChatBot Co-Pilot Umgebung + GenKI-Framework*).
- Fokus auf **kleine, lokale open-weight LLMs** mit **Fine-Tuning** → **Evaluation** weiterer LLMs (Mistral, ...)