

# MULTI-AGENT DESIGN FLOW FOR HIGH-PERFORMANCE SYSTEMS

CHRIS CHENG, PRIYANK KASHYAP, YONGJIN CHOI, JACKSON WEN, NIRJHOR ROUF  
HYBRID CLOUD DIVISION, HEWLETT PACKARD ENTERPRISE, NC STATE



# MOTIVATION

- **Can LLMs be useful for SI/PI and system design?**
  - There has been integration of LLM flows for IC design [1-3] but there has been limited use cases for SI/PI and system design
- **Can LLMs work with strict data restrictions?**
  - Data controls prevent sharing and/or releasing any documentation such as platform design guides (PDGs) externally
- **Can LLMs address productivity bottlenecks?**
  - Internal studies at Nvidia, indicate that LLM-based engineering assistants could alleviate productivity bottlenecks by handling routine tasks that consume substantial engineering time
  - Open source LLMs do not support engineering optimization algos yet

[1] J. Blocklove, S. Garg, R. Karri, and H. Pearce, "Chip-chat: Challenges and opportunities in conversational hardware design," in ACM/IEEE 5th Workshop on Machine Learning for CAD (MLCAD), IE, 2023, pp. 1–6.

[2] K. Chang, K. Wang, N. Yang, et al., "Data is all you need: Finetuning LLMs for chip design via an automated design-data augmentation framework," arXiv preprint arXiv:2403.11202, 2024.

[3] H. Wu, Z. He, X. Zhang, et al., "ChatEDA: A large language model powered autonomous agent for EDA," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2024



# OUR APPROACH - RETRIEVAL AUGMENTED GENERATION

- As the name suggest, RAG enables the LLM to augment its text based on retrieved text
- RAG is low-compute, offline and on-prem solution to provide LLMs with domain specific knowledge [1]
- Can be used to adapt to whatever is present in the database without any additional fine-tuning
  - Code generation
  - Document QA

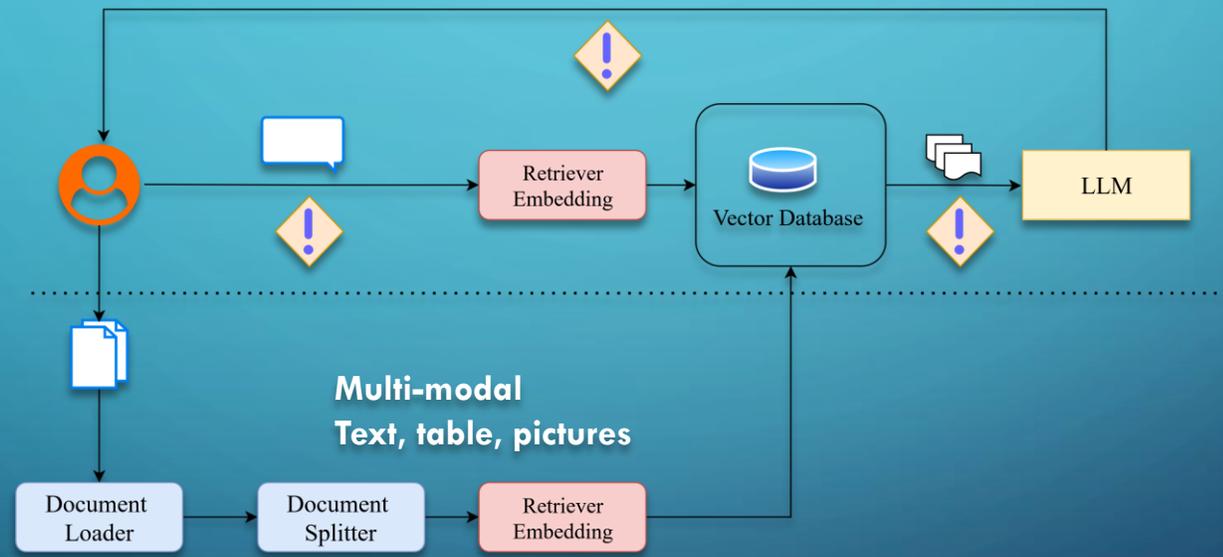
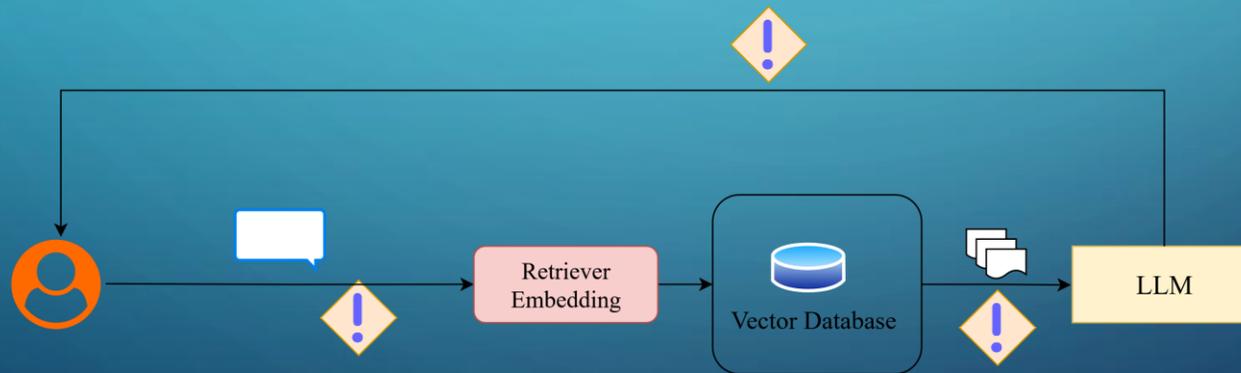


Fig. Naïve RAG pipeline with dedicated flows for document processing and handling user queries

[1] Y. Gao, Y. Xiong, X. Gao, et al., "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2023.

# GUARDRAILS: LLM-AS-A-JUDGE

- LLM-as-a-judge: Use a secondary LLM to evaluate answer and retrieval quality
  - Context relevance: Whether the retrieved documents are relevant to the user query [1]
  - Answer relevance: Whether the answer is relevant to the user's query [1]
- In our work the following is the threshold after experimentation:
  - $threshold = 0.25 \times val_{answer\_rel} + 0.75 \times val_{context\_rel}$
- If the combination of the two is below 0.6, response defaults to
  - *Sorry, there was an error in my response*



# DOCUMENT QA

**Prompt:** What are the standard via dimensions?

**Response:** The term "via dimensions" is not a well-known or widely used term in physics or mathematics. However, I can try to provide some possible interpretations or related concepts:

1. In physics, "via" is not a commonly used term to describe dimensions. The three spatial dimensions we are familiar with are length, width, and height, and the fourth dimension is time. There are also concepts like extra dimensions, which are hypothetical dimensions beyond the three spatial dimensions and one time dimension that we experience.
2. In mathematics, a "via" is not a standard term to describe dimensions. However, in some contexts, "via" can be used as a prefix to describe a path or a route, such as "via negativa" which is a philosophical concept that describes the path of negation or elimination to arrive at a truth.
3. In a more general sense, "via dimensions" could be interpreted as a metaphorical or poetic concept, describing the various ways or paths that one can take to achieve a goal or understand a concept. In this sense, the "via dimensions" would be the different perspectives, approaches, or methods that one can use to arrive at a particular understanding or outcome.

Without more context or information, it is difficult to provide a more specific answer. If you could provide more details or clarify what you mean by "via dimensions," I would be happy to try and assist you further.

(a) No-RAG

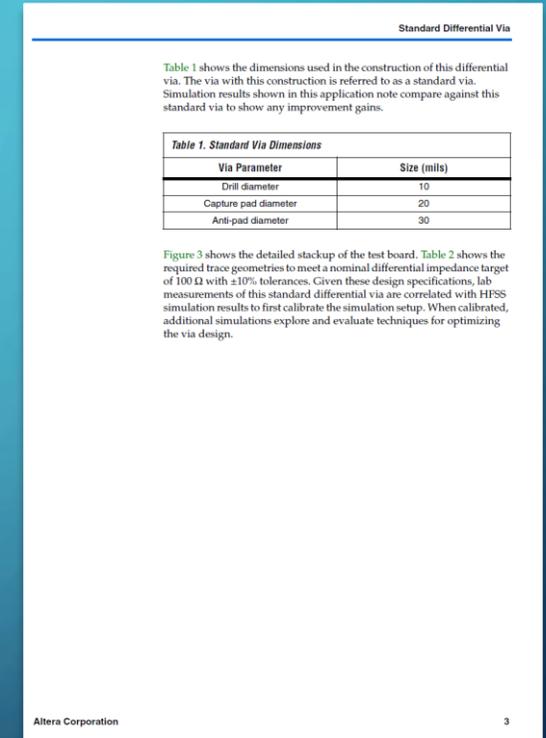
**Prompt:** What are the standard via dimensions?

**Response:** Based on the provided documents, specifically Document 0, Table 1 shows the dimensions used in the construction of the standard differential via. The via dimensions are:

- \* Drill diameter: 10 mils
- \* Capture pad diameter: 20 mils
- \* Anti-pad diameter: 30 mils

These dimensions are referred to as the standard via dimensions.

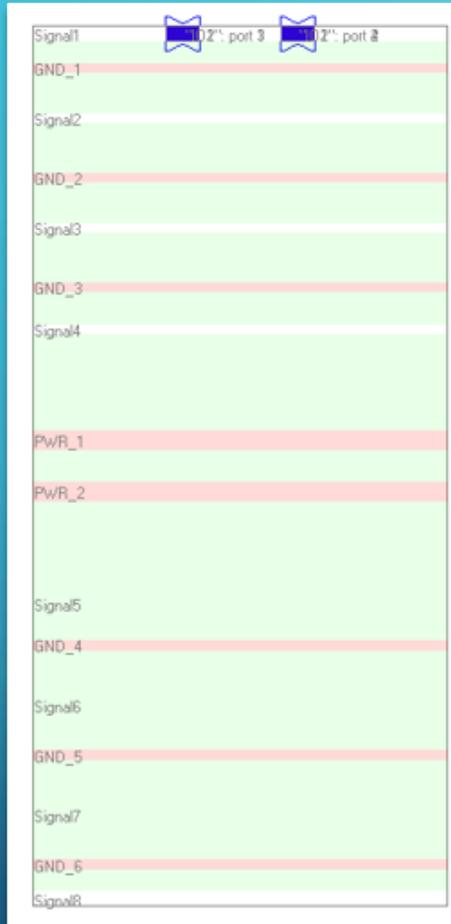
(b) RAG



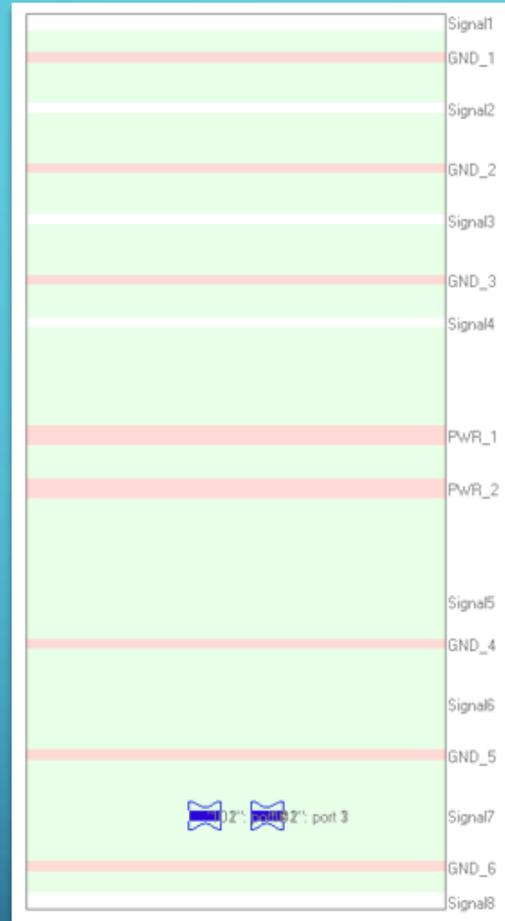
(c) Source Document

Fig. Comparison of the base model performance versus one that has relevant documents provided via RAG. The RAG flow (b) produces succinct response about the via parameters, whereas the no context approach (a) does not provide any specific information about.

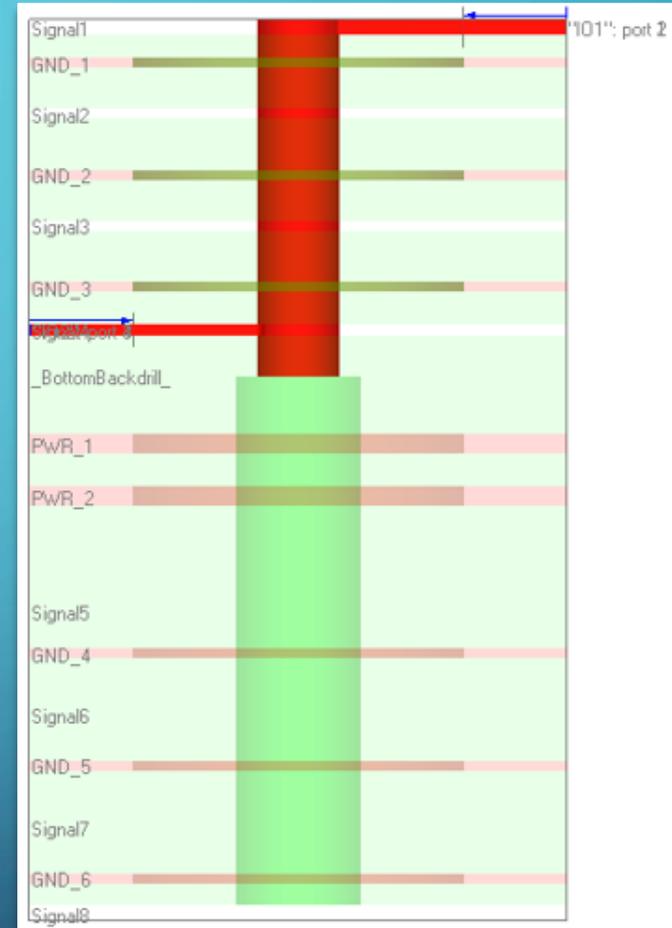
# STRUCTURE GENERATION



(a) Differential Microstrip



(b) Differential Stripline



(c) Differential Via

Fig. Different structures created using code from the LLM using different prompts. The structures from the generated code show the ability for the LLM to perform various tasks

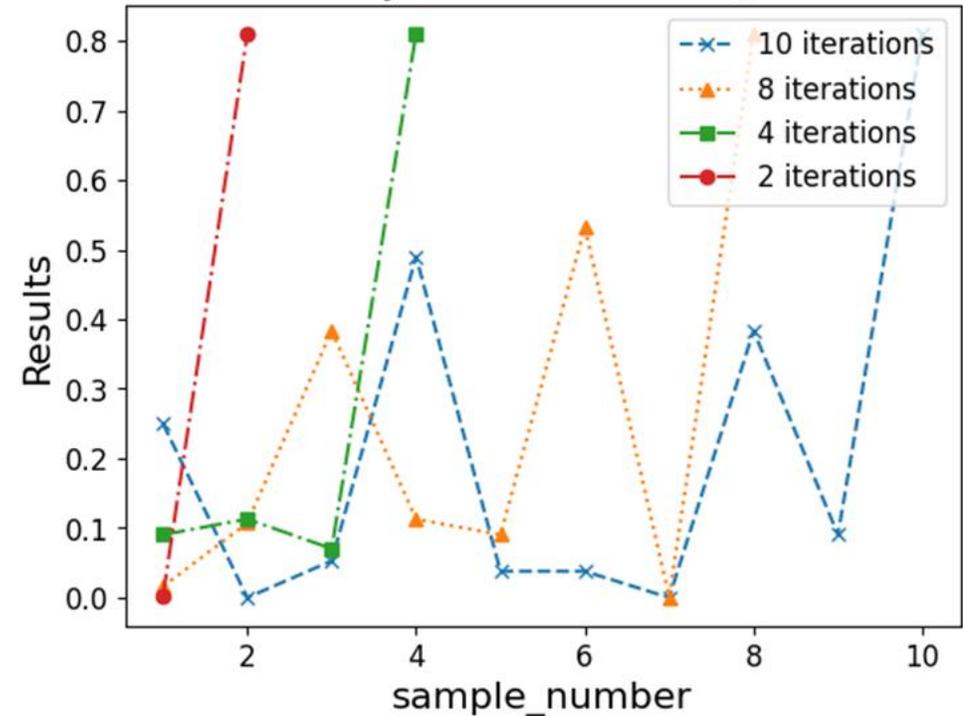
# OPTIMIZATION VIA LLM

- We approached this problem in 2 different ways
  - Optimization by prompting
  - Tool based optimization
- We used Llama-3.1-8B and Llama-3.2-3B for these experiments

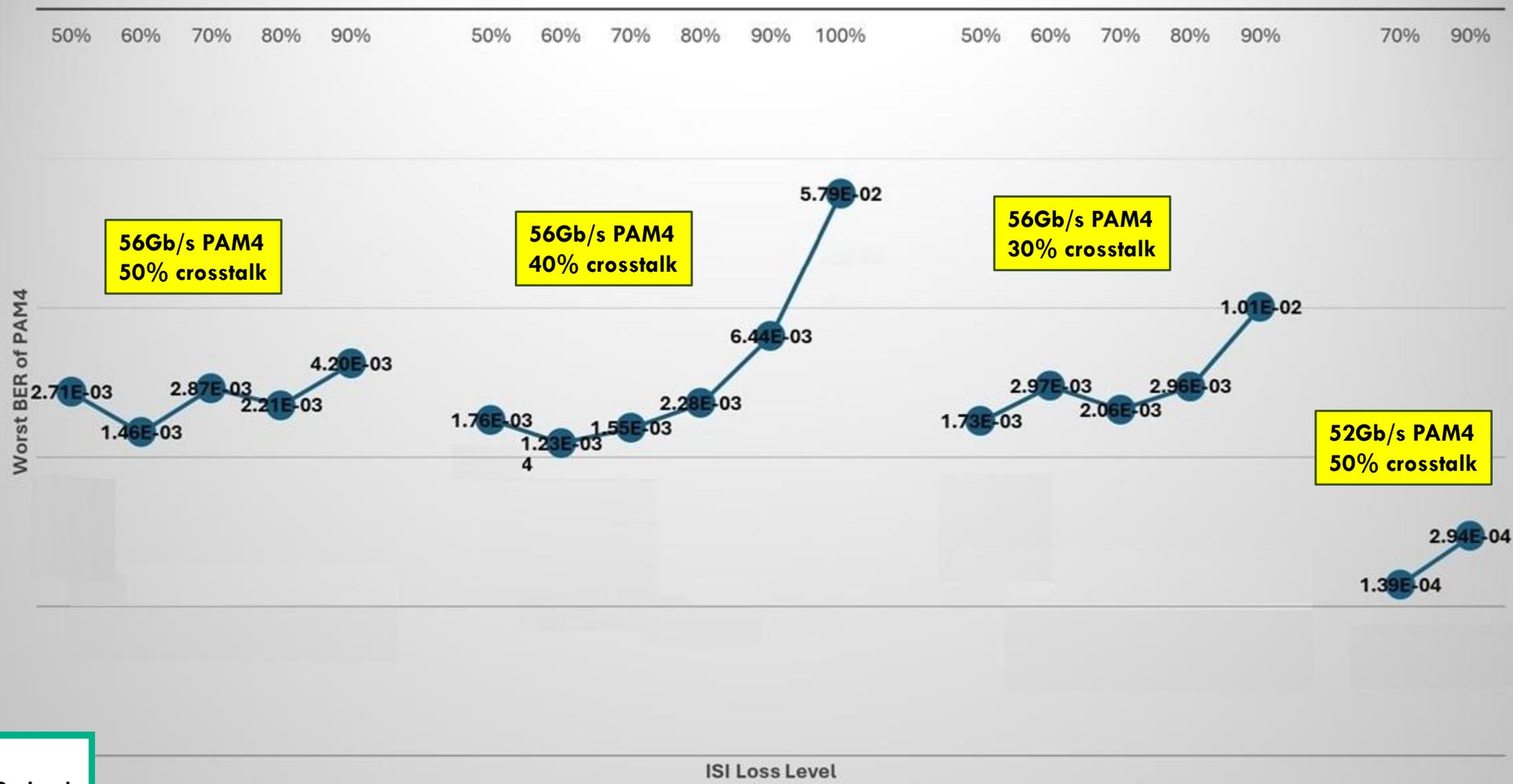
# OPTIMIZATION BY PROMPTING

- The LLM exploited the known results rather than exploring the design space
- LLMs are not good at optimizing for large problems

Evaluation of the objective function @ 5 initial samples



# PAM4 SERDES: AUTO-TUNE VS ML AGENT



**52Gb/s PAM4  
50% crosstalk**

$2.94E-04$   
 $1.39E-04$

# PAM4 SERDES: AUTO-TUNE VS ML AGENT

Deploy

Hello

Base

RAG

Agent ACC

Agent BO

Select Model

meta-llama\_Llama-3.1-8...

Select Tools

run\_serdes\_def... x

optimize\_serdes x

plot\_eye x

eye\_run\_score x

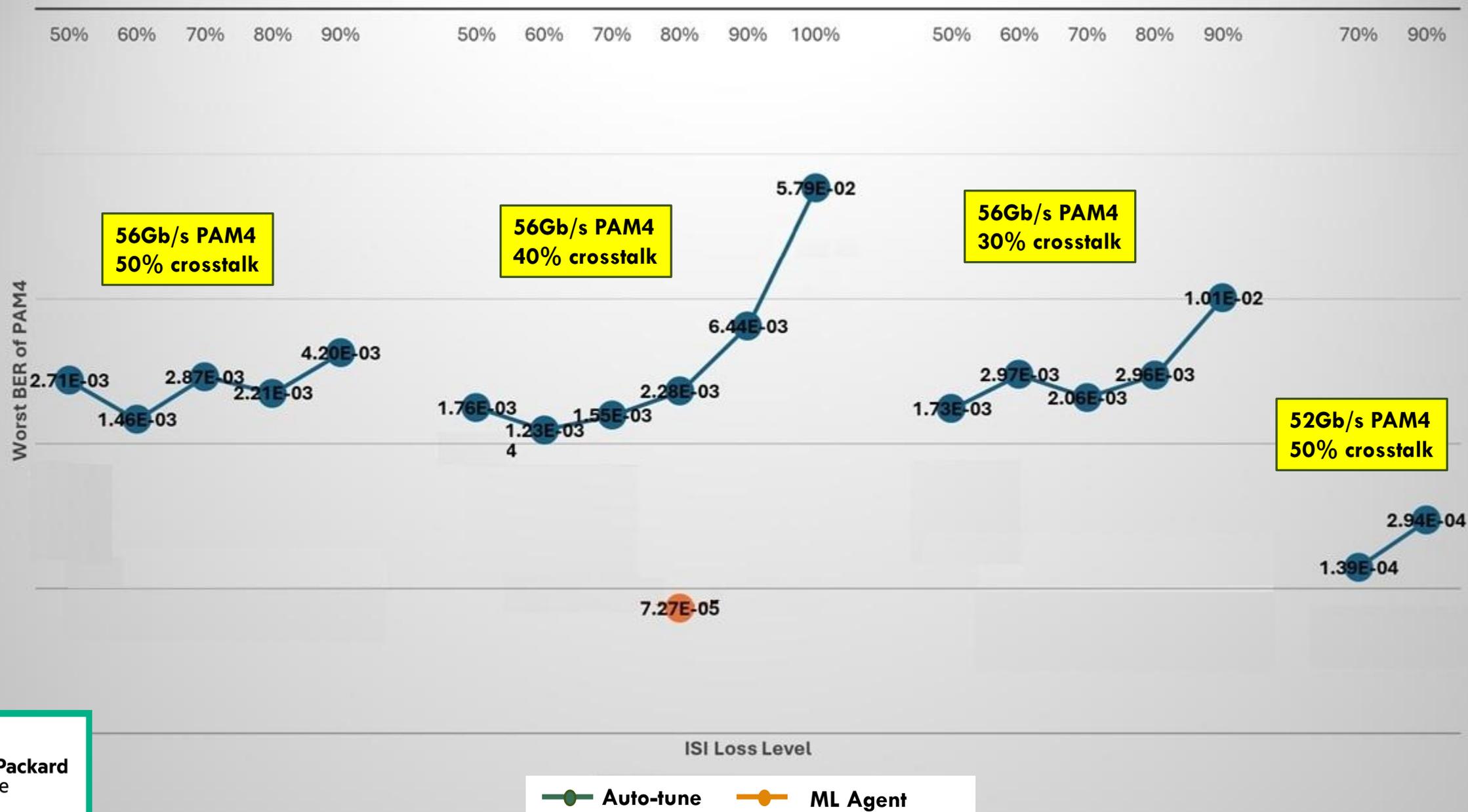
## SerDes Agent

Ask a question...

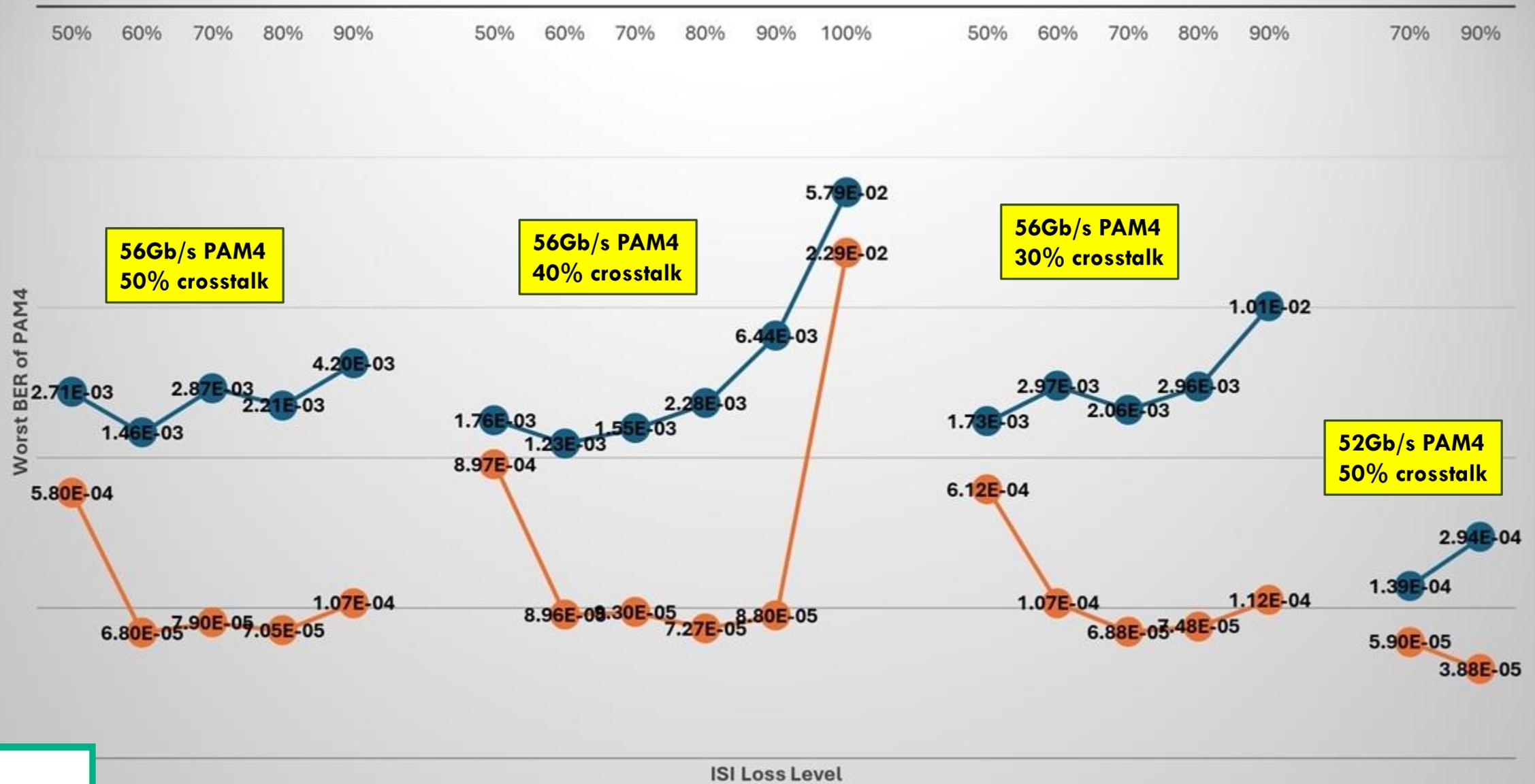
Autotune the SerDes Plot eye diagram Optimize the SerDes

Legal disclaimer

# PAM4 SERDES: AUTO-TUNE VS ML AGENT



# PAM4 SERDES: AUTO-TUNE VS ML AGENT





# PUTTING THEM TOGETHER

## Concept to Design with AI Agent

AI Agent

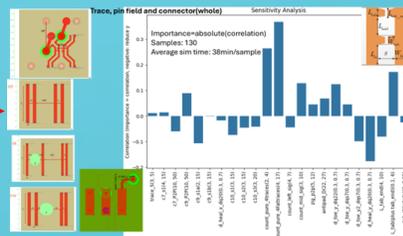


**Llama RAG Flow**

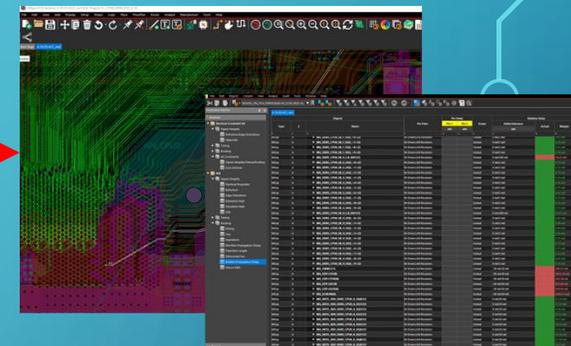
What is the length matching rule between the DQS differential pairs?

0.75 ps

AI Agent

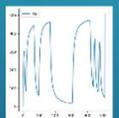
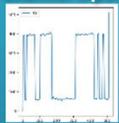


AI Agent

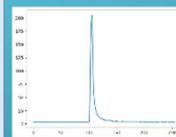
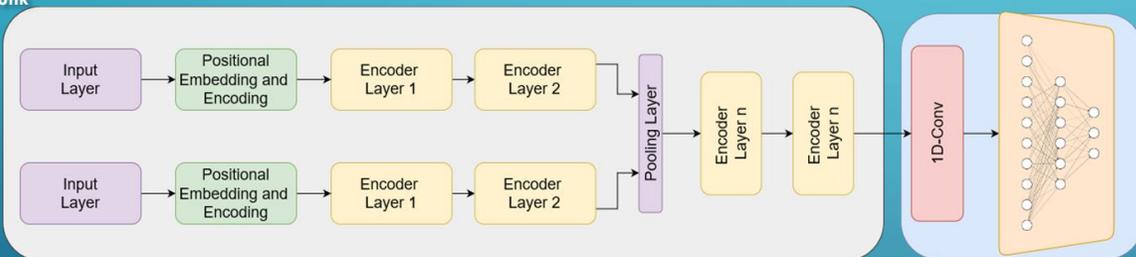


## Test & Validation Phase with Foundational models

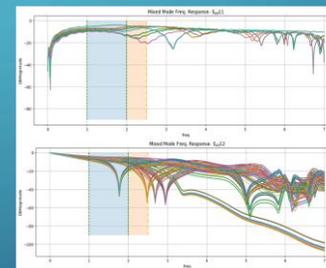
Channel Input Chunk



Channel Output Chunk

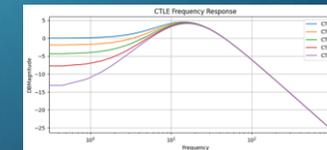


Fine tuning

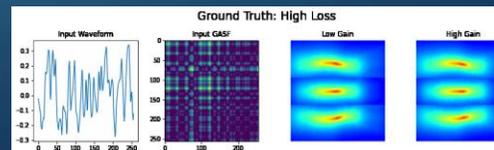
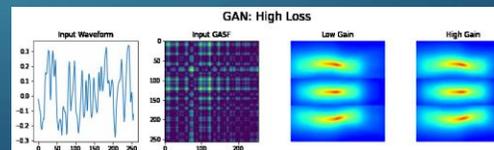
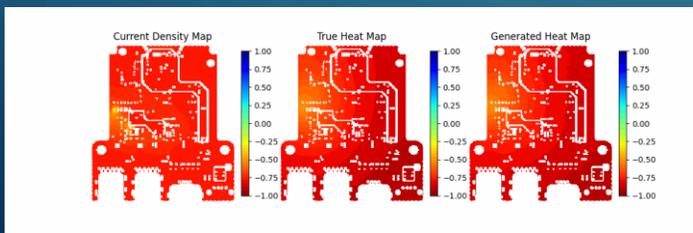


Future-proof

Transfer learning



## Production and deployment with Digital Twins



**AI companies  
want you to think  
you are this guy**



**LLM**

Thank you



**Hewlett Packard**  
Enterprise



BACKUP

# RAG – DOCUMENT FLOW

- Load the documents using local libraries to pre-process the documents
  - Neural Optical Understanding for Academic Documents (Nougat) [1]
  - Converts text and tables into markdown format for later use
  - Needed to integrate PDGs
- Split the documents into smaller chunks
  - Chunk size is a hyper-parameter
    - For PDGs we limit to a single page
    - For non-PDGs we chunk based on title
- Embed the chunks using an embedding model
  - all-MiniLM-L6-v2 [2]
- Store the embeddings in a FAISS vector database [3]
  - Built in a similarity search to find relevant documents



Fig. Document processing flow for RAG database

[1] L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic, Nougat: Neural optical understanding for academic documents, 2023. arXiv: 2308.13418 [cs.LG].

[2] Available: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>.

[3] L. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2019.

# RAG – USER FLOW

- User query goes through the same embedding as the documents
- Using FAISS DB perform a similarity search in the database and return *top-n* document chunks
- Add the retrieved chunks to the query to enable the LLM to have domain knowledge
- The additional context does not make the hallucination problem go away

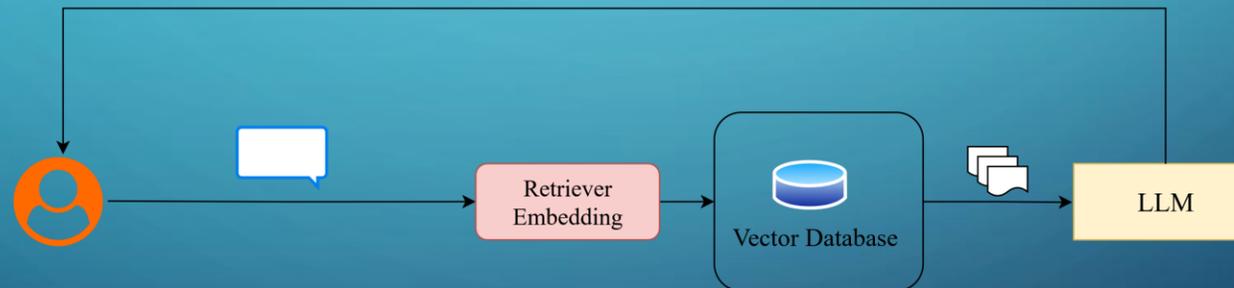


Fig. User query pipeline through a database and LLM for QA and code generation