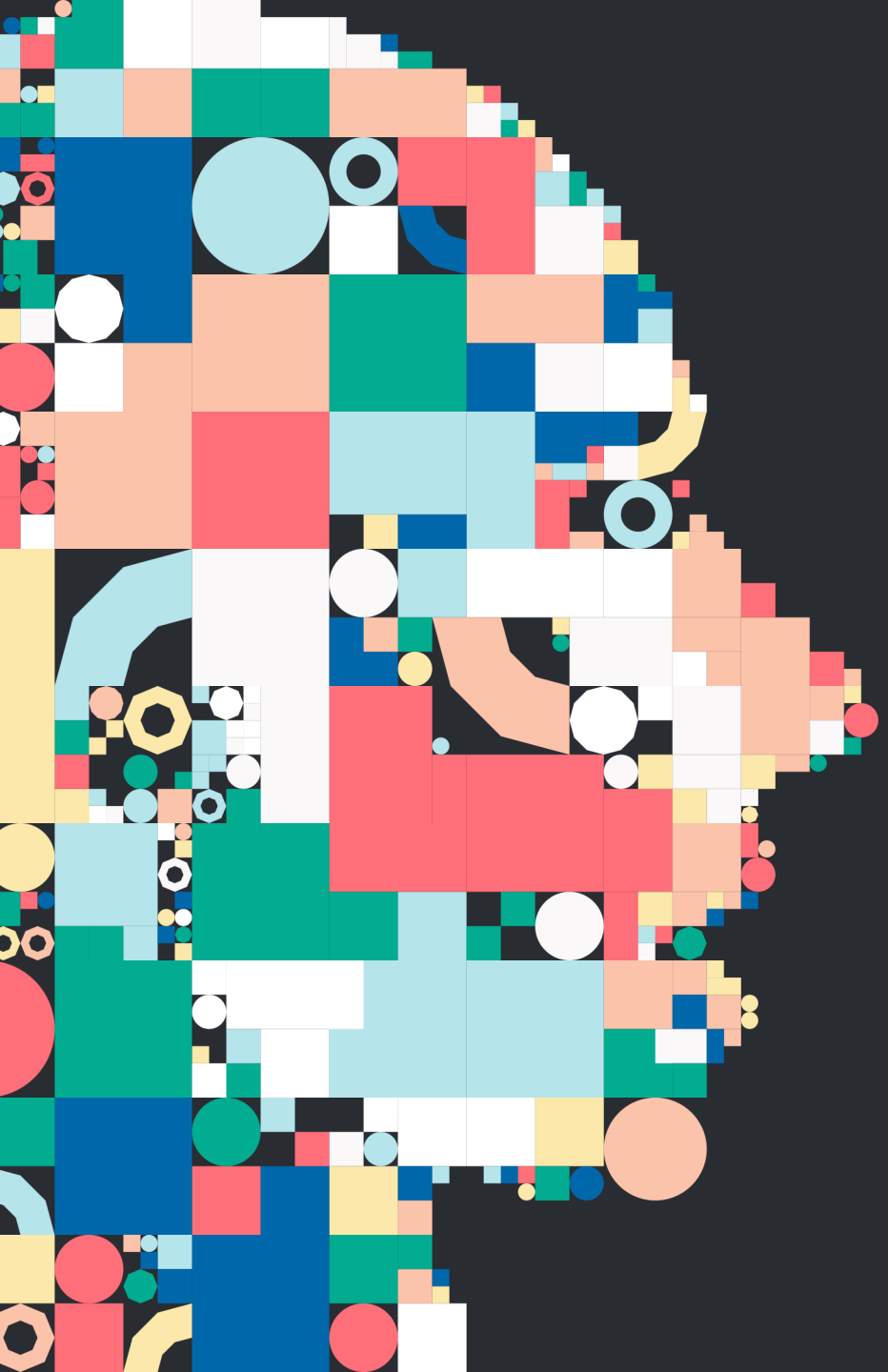


WELCOME TO THE GENERATIVE AI MEETUP NYC

5:30 - 5:45

GPT4ALL: OPEN SOURCE, ON-EDGE LARGE LANGUAGE MODELS

Andriy Mulyar, Founder & CTO, Nomic AI

5:45 – 6:00

UNDERSTANDING THE LANDSCAPE OF LARGE LANGUAGE MODELS

Estella Xin, Software Engineer, ML Workflow Team, W&B

6:00 – 6:10

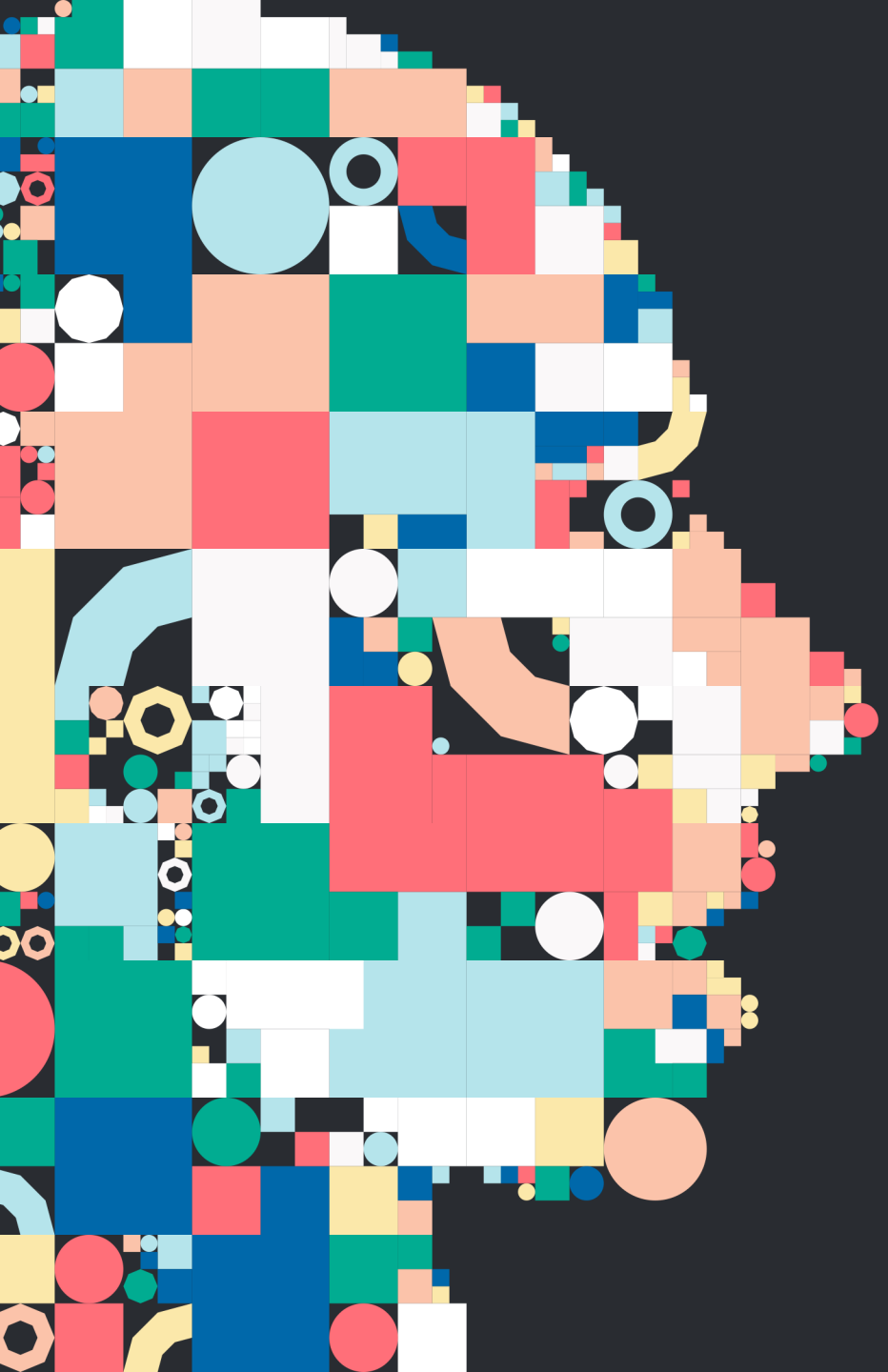
WHY IPUS FOR GENERATIVE AI MODELS?

Helen Byrne, VP Solution Architects, Graphcore

6:10 – 6:20

DEMO: DEPLOYING YOUR GENERATIVE AI APPS IN THE CLOUD

Zack Zweig, Product Manager, Paperspace



WELCOME TO THE GENERATIVE AI MEETUP NYC

6:20 – 6:30 **10 MIN BREAK**

6:30 – 7:00 **FIRESIDE CHAT + Q&A**
Dillon Erb, CEO, Paperspace
Andriy Mulyar, Founder & CTO, Nomic AI
Yuliya Shcherbachova, Director Revenue Enablement, W&B
Zack Zweig, Product Manager, Paperspace
with Helen Byrne

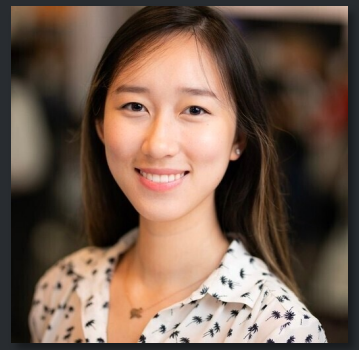
7:00 – 8:30 **NETWORKING DRINKS**

ANDRIY MULYAR

FOUNDER & CTO

NOMIC AI

GPT4All: Open Source, On-Edge Large Language Models



ESTELLA HIN

SOFTWARE ENGINEER, ML WORKFLOW TEAM

WEIGHTS & BIASES

Demo: Optimize LLM training with W&B

W&B

Estella Xin

Demo: Optimize LLM training with W&B

W&B for LLMs | ML Practitioner Types

1. **LLM Creators: train LLMs from scratch**
2. **LLM Fine-Tuners: adapt existing LLMs to specific tasks**
3. **LLM Prompt Engineers: use LLM as API service with prompt tuning**

W&B for LLMs | ML Practitioner Types

1. **LLM Creators: train LLMs from scratch**
2. **LLM Fine-Tuners: adapt existing LLMs to specific tasks**
3. **LLM Prompt Engineers: use LLM as API service with prompt tuning**

LangChain

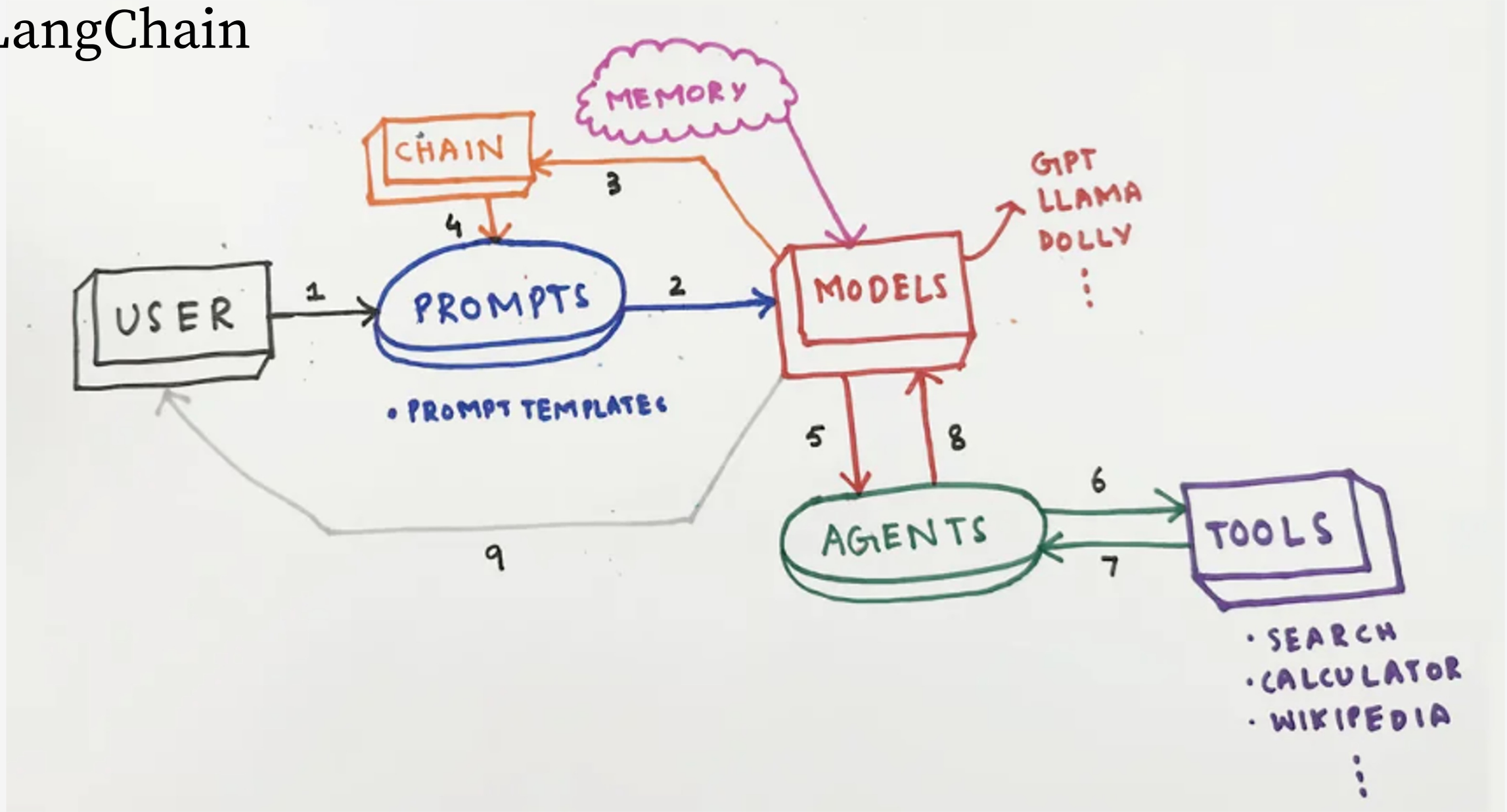


Image by Brilliant Bytes

W&B

W&B Prompts, our LLMOps toolset!

Use a human readable question to generate a sql query

Input:

- Find the ID of the first product we sent a customer
- Sql Schema of database

Output:

```
  SELECT L_PARTKEY
FROM TPCH_SF1.LINEITEM
INNER JOIN TPCH_SF1.ORDERS ON L_ORDERKEY = O_ORDERKEY
GROUP BY O_CUSTKEY, L_PARTKEY
ORDER BY O_CUSTKEY, L_LINENUMBER;
```

We use langchain to make a set of steps:

1) Generate SQL
2) Execute SQL

Enable Wandb tracing

```
from langchain.callbacks import wandb_tracer

wandb_args = {"project": config.WANDB_PROJECT, "entity": config.WANDB_ENTITY}

wandb_tracer.watch(wandb_args)
```

✓ 0.6s

```
wandb: W&B Run initialized. View LangChain logs in W&B at https://wandb.ai/prompt-eng/mt-pocono/runs/bsmofqh3.
```

```python
# Chain 1: Generate SQL query from a user question
llm = OpenAI(openai_api_key=config.OPENAI_API_KEY,
             model_name = "text-davinci-003",
             temperature=0,
             verbose=True)

template = f"Here is a snowflake database schema: {{schema_str}}.{{question}}"

generate_sql_chain = LLMChain(
    llm=llm,
    prompt=PromptTemplate(input_variables=["schema_str", "question"], template=template),
    output_key="sql",
    verbose=True)

# Chain 2: Run the SQL query
def run_sql(inputs: dict) -> dict:
    return {"sql_result": sql_conn(inputs["sql"])}

run_sql_chain = TransformChain(
    input_variables=["sql"],
    output_variables=["sql_result"],
    transform=run_sql,
    verbose=True)

# Wrap the two chains into a SequentialChain
sql_chain = SequentialChain(
    chains=[generate_sql_chain, run_sql_chain],
    input_variables=["schema_str", "question"],
    output_variables=["sql_result", "sql"],
    verbose=True)
```

Lets go look at the Wandb UI!

Can the LLM correct itself?

New set of steps:

1) Generate SQL
2) **Ask LLM to Correct SQL**
3) Execute SQL

```
# Chain Step 2: Cleanup and Format SQL: {raw_sql} -> {clean_sql}
clean_sql_template = f"""Please correct any syntax errors in the following SQL and format it nicely:

{{raw_sql}}

Correct SQL:"""

generate_sql_chain = LLMChain(
    llm=llm,
    prompt=PromptTemplate(input_variables=["schema_str", "question"], template=template),
    output_key="raw_sql",
    verbose=True)

clean_sql_chain = LLMChain(
    llm=llm,
    prompt=PromptTemplate(input_variables=["raw_sql"], template=clean_sql_template),
    output_key="sql",
    verbose=True)

run_sql_chain = TransformChain(
    input_variables=["sql"],
    output_variables=["sql_result"],
    transform=run_sql,
    verbose=True)

# Wrap the two chains into a SequentialChain
sql_chain = SequentialChain(
    chains=[generate_sql_chain, clean_sql_chain, run_sql_chain],
    input_variables=["schema_str", "question"],
    output_variables=["sql_result", "sql"],
    verbose=True)
```

Lets go look at the Wandb UI!

Close Wandb tracing

```
wandb_tracer.finish()
✓ 5.9s
wandb: All files uploaded. View LangChain logs in W&B at https://wandb.ai/prompt-eng/
```

How to we explore large sets of generations?

```
runs.map((row) => row.history["langchain_trace"]).dropna.concat.dropna
```

All Traces

	Success	Timestamp	Input	Output	Chain	Error
1	True	2023-04-20 13:14:47	**0.question:** What is my best performing region	**0.sql_result:** EUROPE,44032702326.2956	SequentialChain(LLMChain(OpenAI), TransformChain)	-
2	True	2023-04-20 13:15:02	**0.question:** Find the id of the first product we sent every customer	**0.sql_result:** 64762,150129,91322,85546,189101,47242,56410,63064,159338,37152,175[illegible]	SequentialChain(LLMChain(OpenAI), TransformChain)	-
3	True	2023-04-20 13:15:14	**0.question:** how many orders have been made	**0.sql_result:** 1500000	SequentialChain(LLMChain(OpenAI), TransformChain)	-
4	True	2023-04-20 13:15:19	**0.question:** how many orders have been made in the last 30 days	**0.sql_result:** 0	SequentialChain(LLMChain(OpenAI), TransformChain)	-
5	True	2023-04-20 13:15:25	**0.question:** how many products are there	**0.sql_result:** 200000	SequentialChain(LLMChain(OpenAI), TransformChain)	-
6	True	2023-04-20 13:15:29	**0.question:** how many customers are there	**0.sql_result:** 150000	SequentialChain(LLMChain(OpenAI), TransformChain)	-
7	True	2023-04-20 13:15:31	**0.question:** Get my last 10 orders	**0.sql_result:** 6000000,110063,O,37625.29,841449600000,2-HIGH,Clerk#000000411,0,ess pinto beans boost slyly regular accounts[illegible]	SequentialChain(LLMChain(OpenAI), TransformChain)	-
8	True	2023-04-20 13:15:35	**0.question:** what is each customers total spend	**0.sql_result:** 1,Customer#000000001,564945.3484,2,Customer#000000002,975007.7642[illegible]	SequentialChain(LLMChain(OpenAI), TransformChain)	-

← ‹ 1 - 8 of 101 › → Export as CSV Columns... Reset Table

Trace Timeline (#101) Model Architecture (ID: 77ac02bb70c10b7f)

1: SequentialChain 8865ms

2: LLMChain 7044ms

3: OpenAI 7044ms

4: TransformChain 18

SequentialChain 8865ms

Result Set 1

Inputs

question Find the top 1 customer who has spent the most money

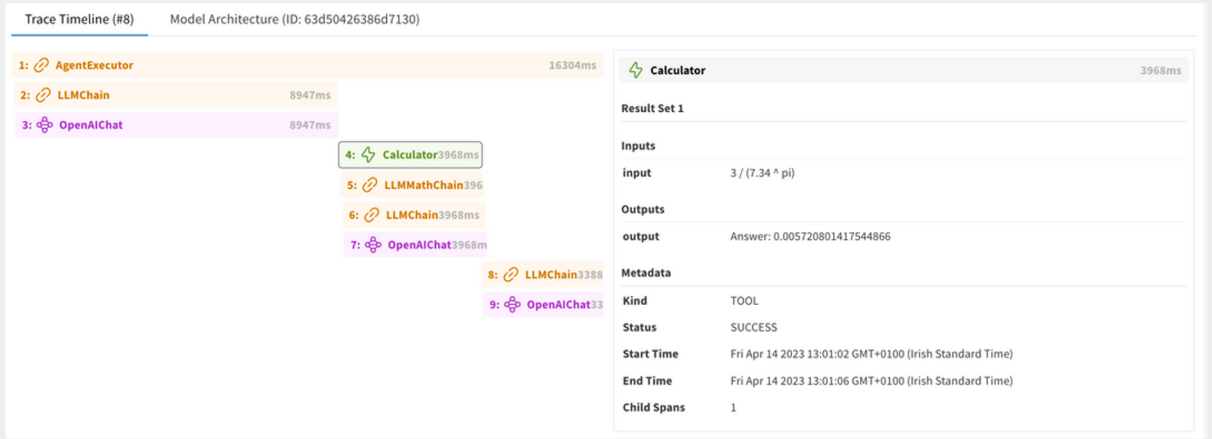
schema_str
```
[{"table": "TPCH_SF1.CUSTOMER", "C_CUSTKEY": "
{\"type\":\"FIXED\",\"precision\":38,\"scale\":0,\"nullable\":false}", "C_NAME": "
{\"type\":\"TEXT\",\"length\":25,\"byteLength\":100,\"nullable\":false,\"fixed\":false}", "C_ADDRESS": "
{\"type\":\"TEXT\",\"length\":40,\"byteLength\":160,\"nullable\":false,\"fixed\":false}", "C_NATIONKEY": "
{\"type\":\"FIXED\",\"precision\":38,\"scale\":0,\"nullable\":false}", "C_PHONE": "
```

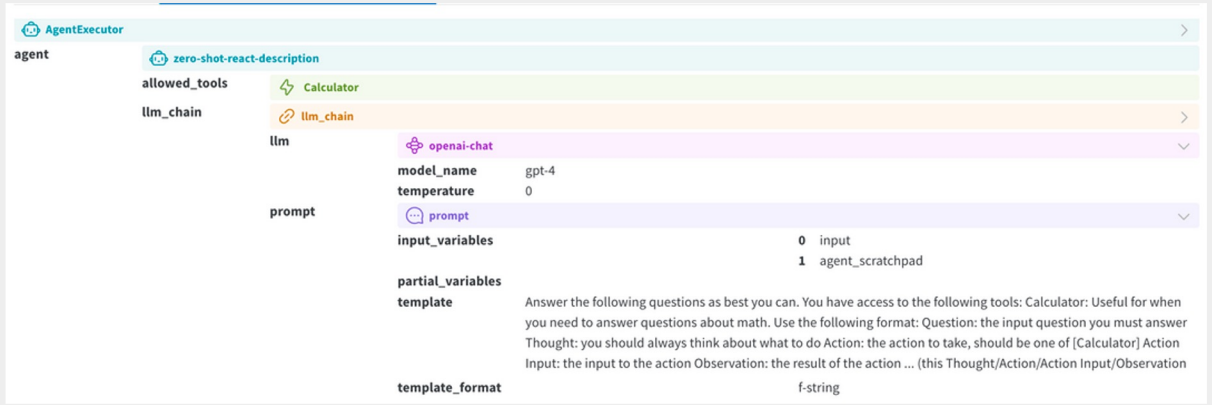
Outputs

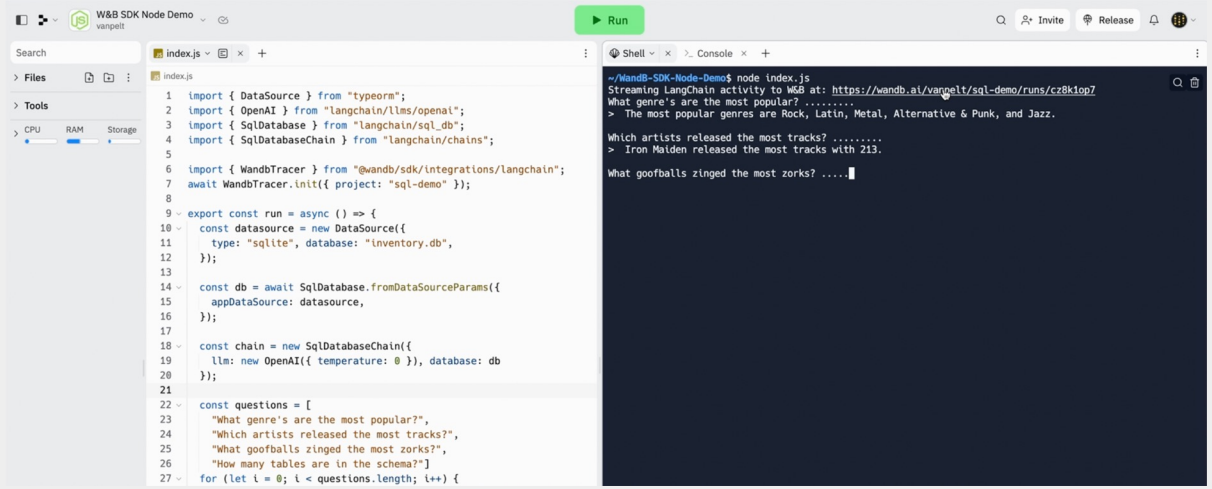
Close

W&B Prompts

- Records execution traces, model topologies, and all of the interactions
- Review and debug errors
- Glean insights about model behavior
- Share learnings & advancements with colleagues







W&B

Get Started with W&B Prompts!

wandb.me/llms
wandb.me/courses

W&B

Thank You!

Questions Index

- Wandb documentation on trace: https://docs.wandb.ai/guides/prompts/quickstart
- Bigger example: https://wandb.ai/prompt-eng/prompts-quickstart?workspace=user-morg

Tables 1

```
runs.map((row) => row.history["langchain_trace"]).dropna.concat.dropna
```

All Traces

	Succe	Timestamp	Input	Output
323	True	2023-06-2 00:18:14	**0.question:** Is it possible to provide any options while using the `wandb offline` command?	**0.text:** The WANDBOT_RESPONSE is accurate. It correctly states that it is possible to provide options while using the `wandb offline` command a
322	True	2023-06-2 00:15:30	**0.question:** What are the deployment options for W&B?	**0.text:** The WANDBOT_RESPONSE accurately states that the documenta does not provide information on deployment options for W&B. Therefor
321	True	2023-06-2 00:14:24	**0.question:** What is the benefit of using W&B with TensorBoard?	**0.answer:** Using W&B with TensorBoard has several benefits, including:
320	True	2023-06-2 00:07:25	**0.question:** Is the JavaScript integration of wandb still in Beta?	**0.answer:** I'm sorry, I don't have that information. The provided context does not mention anything about the current status of the JavaScript
319	True	2023-06-1 23:58:33	**0.question:** Can you use the `WandBLogger` class with the `EarlyStopping` callback in Composer?	**0.answer:** I'm not sure. The documentation for wandb does not mentio the `WandBLogger` class or the `EarlyStopping` callback in Composer

← ‹ 1 - 5 of 323 › →

Trace Timeline (#322)

LLMChain 3945ms

2: ChatOpenAI 3944ms

answer focuses on how to integrate W&B into your m features for experimentation, logging, and vis

source_documents {"source_doc_0": {"page_content": "A full example is available here.\n\nHow is W&B different from TensorBoard?\n\nWhen the cofounders started working on W&B, they were inspired to build a tool for the frustrated TensorBoard users at OpenAI. Here are a few things we've focused on improving:\n\nReproduce models: Weights & Biases is good for experimentation,

Outputs

String preview

{"source_doc_0": {"page_content": "A full example is available here.\n\nHow is W&B different from TensorBoard?\n\nWhen the cofounders started working on W&B, they were inspired to build a tool for the frustrated TensorBoard users at OpenAI. Here are a few things we've focused on improving:\n\nReproduce models: Weights & Biases is good for experimentation, exploration, and reproducing models later. We capture not just the metrics, but also the hyperparameters and version of the code, and we can save your version-control status and

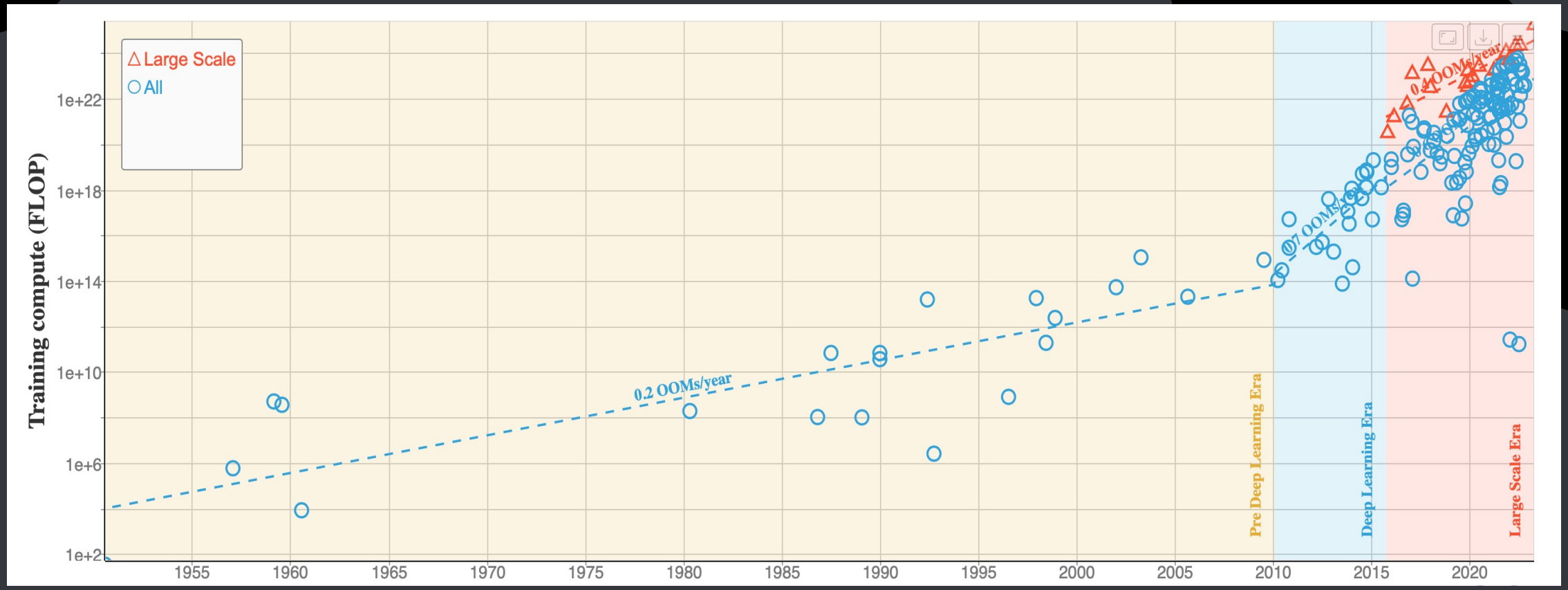
HELEN BYRNE

VP SOLUTION ARCHITECTS

GRAPHCORE

Why IPUs for Generative AI Models?

A NEW ERA OF COMPUTE...



LIMITED CAPACITY OF GPUS

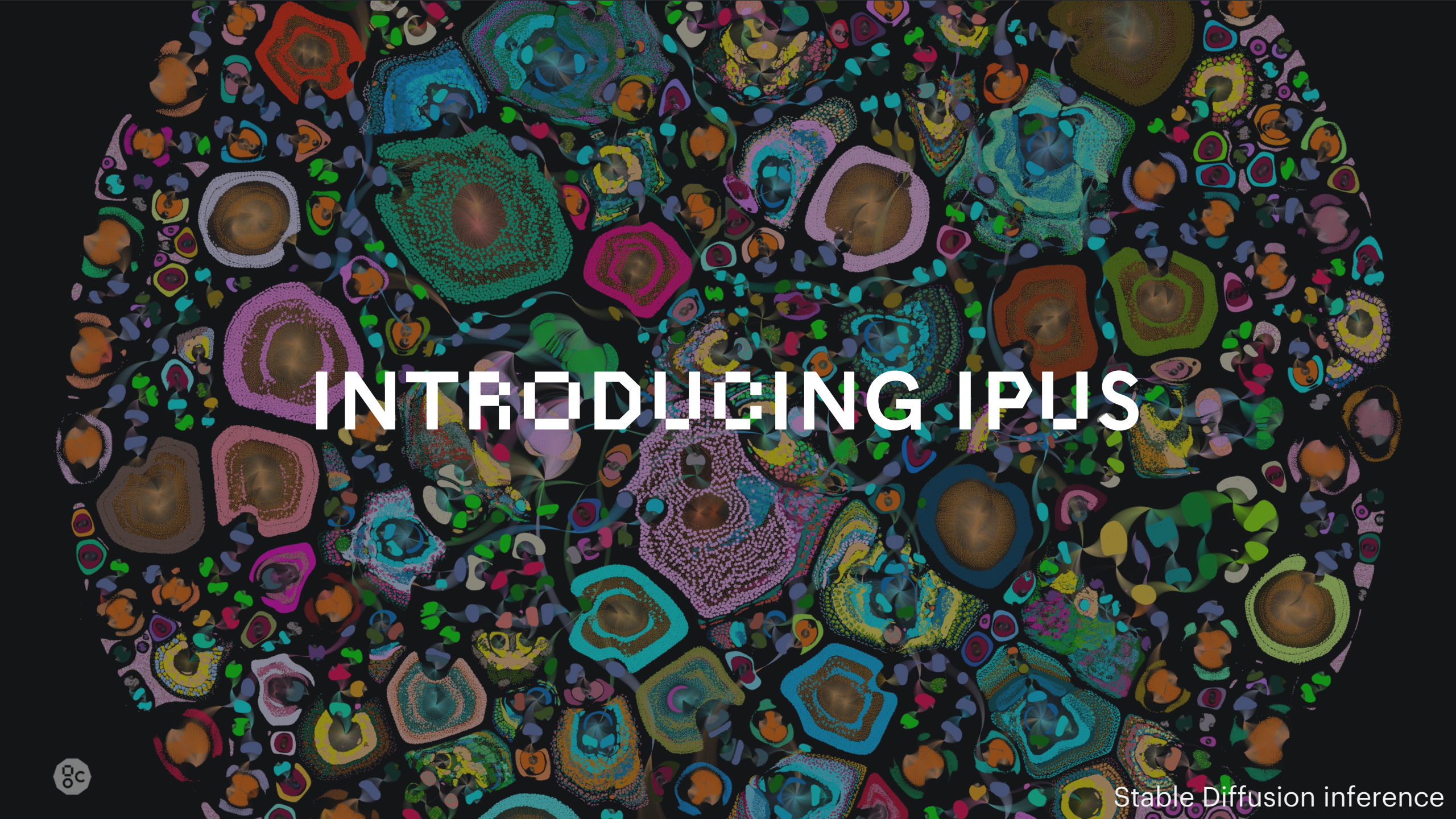
"OpenAI is extremely GPU-limited and this is delaying a lot of their short-term plans."

"The biggest customer complaint was about the reliability and speed of the API... most of the issue was a result of GPU shortages."

"The finetuning API is also currently bottlenecked by GPU availability."

"Dedicated capacity offering is limited by GPU availability."

Humanloop



Stable Diffusion inference


GRAPHCORE
115-0104
N9T239.00
CC90
2109

IPU-Tiles™

1472 independent IPU-Tiles™ each with an IPU-Core™ and In-Processor-Memory™

IPU-Core™

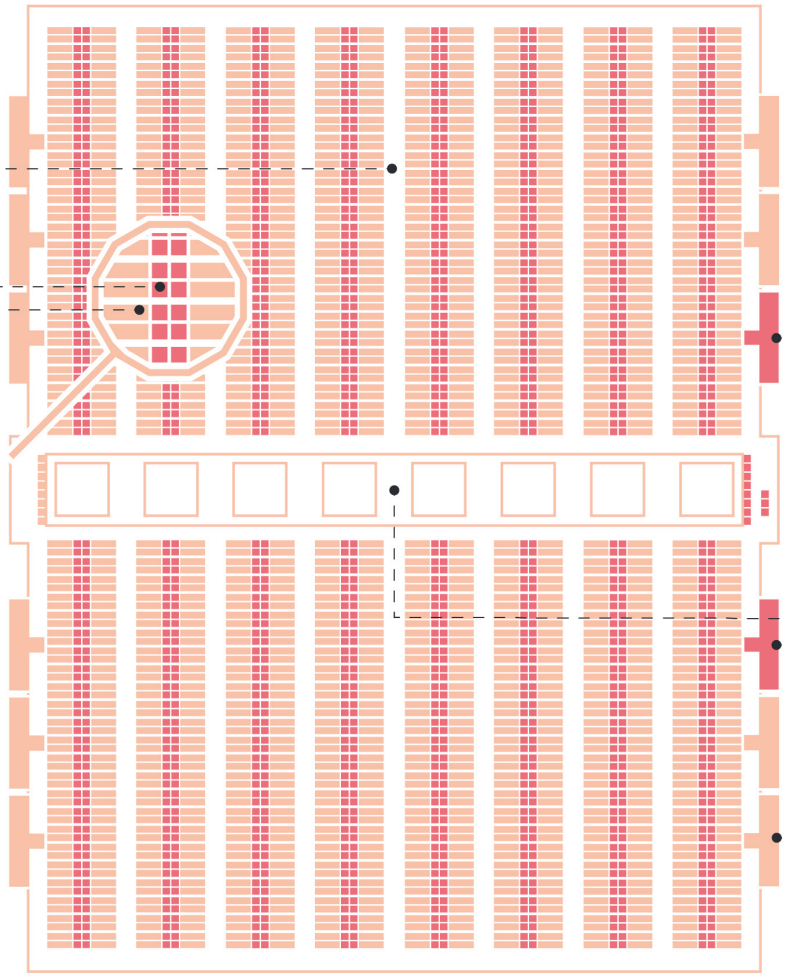
1472 independent IPU-Core™

8832 independent program threads executing in parallel

In-Processor-Memory™

900MB In-Processor-Memory™ per IPU

65TB/s memory bandwidth per IPU



IPU-Exchange™

11 TB/s all to all IPU-Exchange™
Non-blocking, any communication pattern

PCIe

PCI Gen4 x16
64 GB/s bidirectional bandwidth to host

IPU-Links™

10 x IPU-Links,
320GB/s chip to chip bandwidth

ALEPH ALPHA BOT

Click Lumi to open brain.

Settings

Choose available model:

luminous-sparse-graphcore

Choose mode:

information

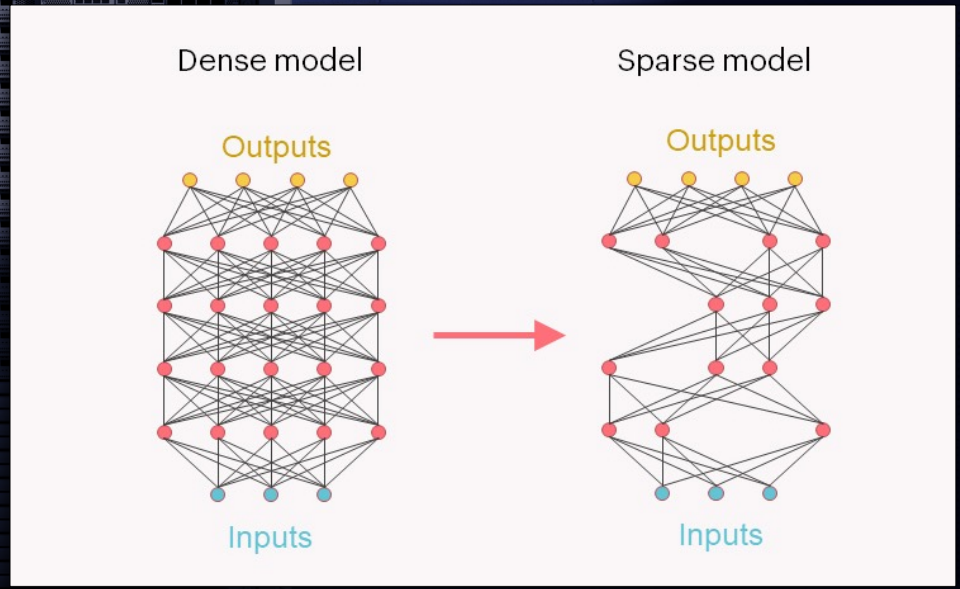
LUMI

Hello! I am Lumi. Ask me questions!

Enter message

Send

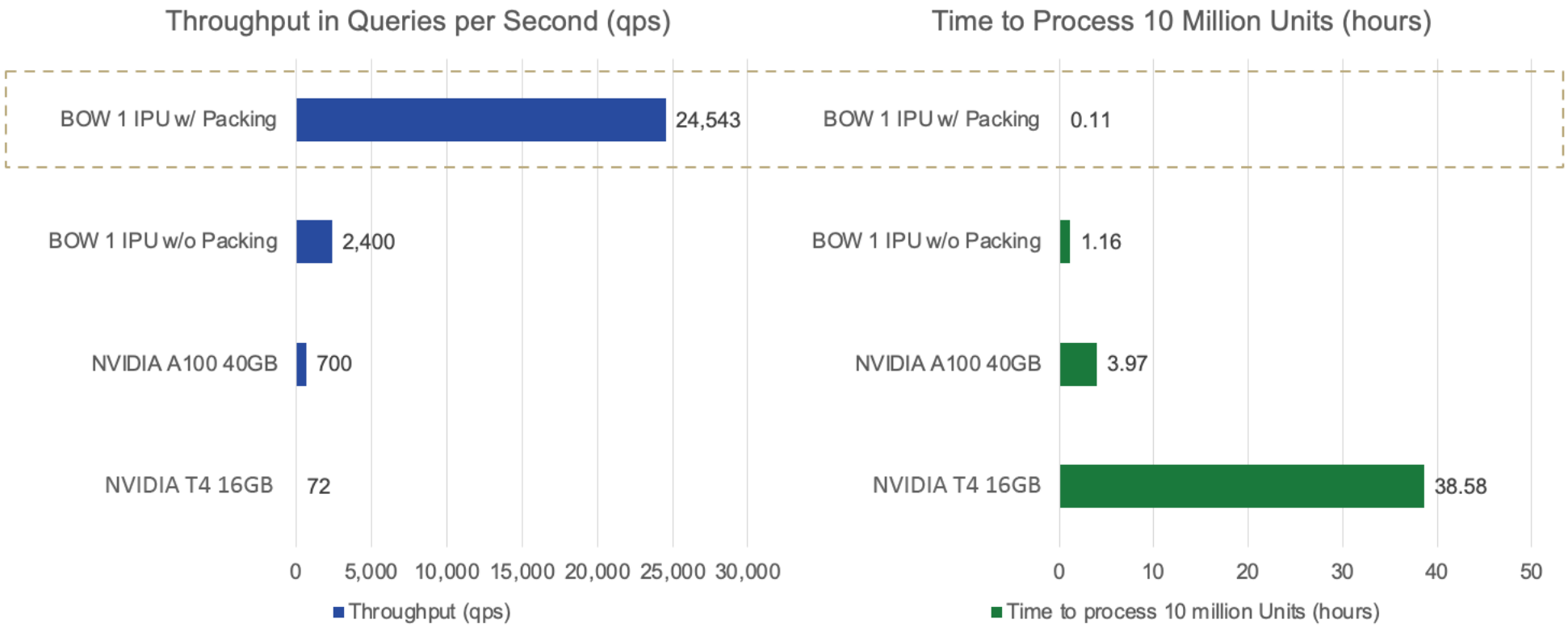
ALEPH ALPHA



Accelerator Performance

Nvidia GPU vs Graphcore IPU

pienso



TRY OUR GENAI & LLM MODELS IN THE CLOUD

Image-to-Image Generation
on IPU with ***Stable Diffusion***

Run on Gradient

Text-to-Image Generation
on IPU with ***Stable Diffusion***

Run on Gradient

Text Guided In-Painting
on IPU with ***Stable Diffusion***

Run on Gradient

Fine-tune ***MT5***
on IPU

Run on Gradient

Inference on text prompts on
IPU with ***Dolly 2.0***

Run on Gradient

Chatbot on IPU with
OpenAssistant Pythia 12B

Run on Gradient

Few-shot learning on IPU with
Flan T5-Large/XL

Run on Gradient

Zero-Shot Text Classification
on IPUs using ***MT5-Large***

Run on Gradient

Fine-tune ***GPT-J 6B*** for
text entailment on IPU

Run on Gradient

Text generation on IPU
using ***GPT-J 6B***

Run on Gradient

Finetune ***BART***
on IPU

Run on Gradient

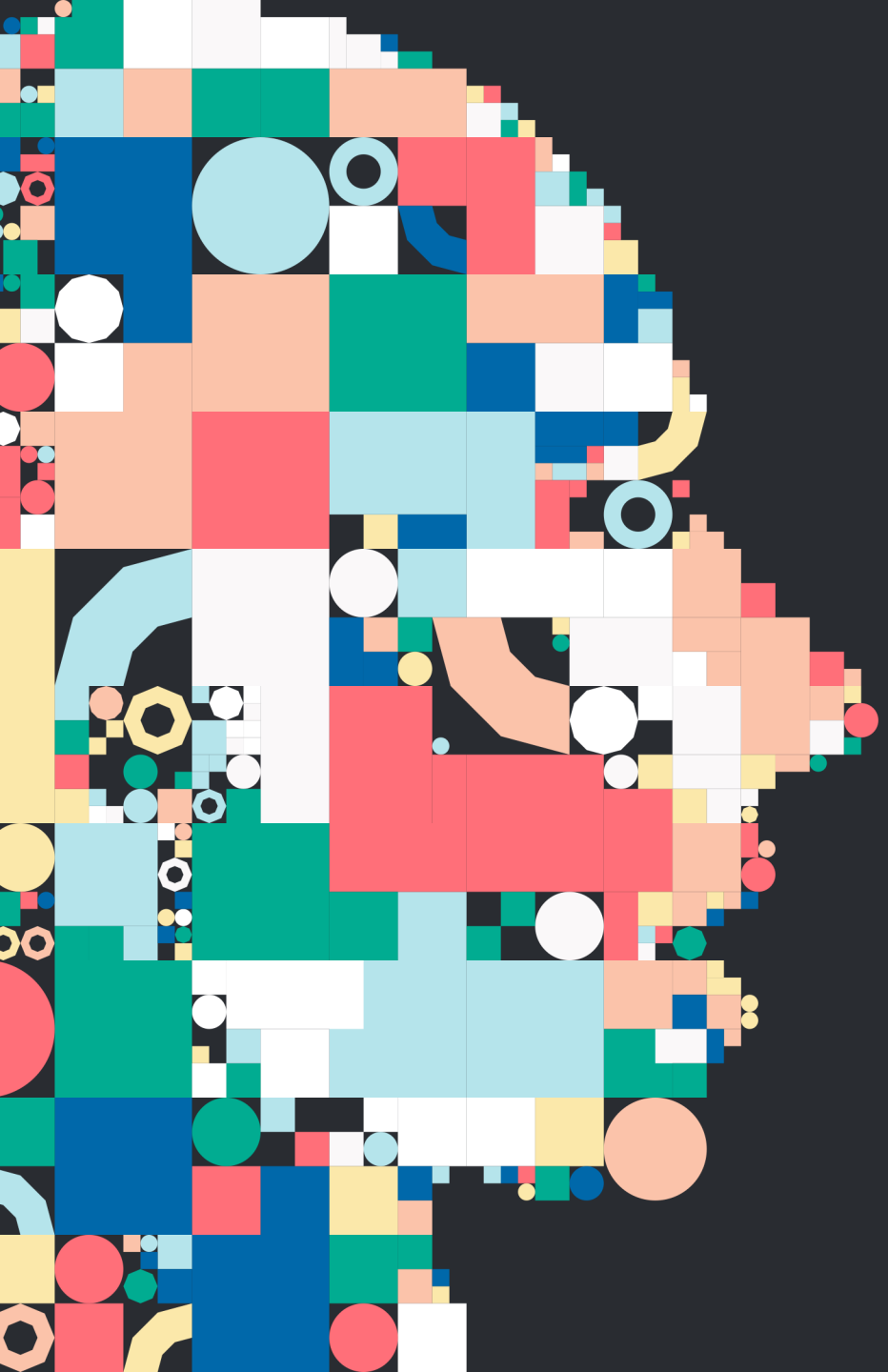
ASR on IPUs
using ***Whisper***

Run on Gradient

ipu.dev/notebooks

WHY IPUS FOR GENERATIVE AI MODELS?

1. A NEW ERA OF COMPUTE... IPUS DESIGNED FOR AI
2. CAPACITY WITH OUR CLOUD PARTNER *Paperspace*
3. LATENCY, THROUGHPUT & COST ADVANTAGE
4. GET STARTED WITH EASY-TO-RUN EXAMPLES
5. SOLUTION ARCHITECTS TEAM READY TO SUPPORT



ANNOUNCING

GRAPHCORE AI STARTUP PROGRAMME

Supercharge your startup with IPU compute

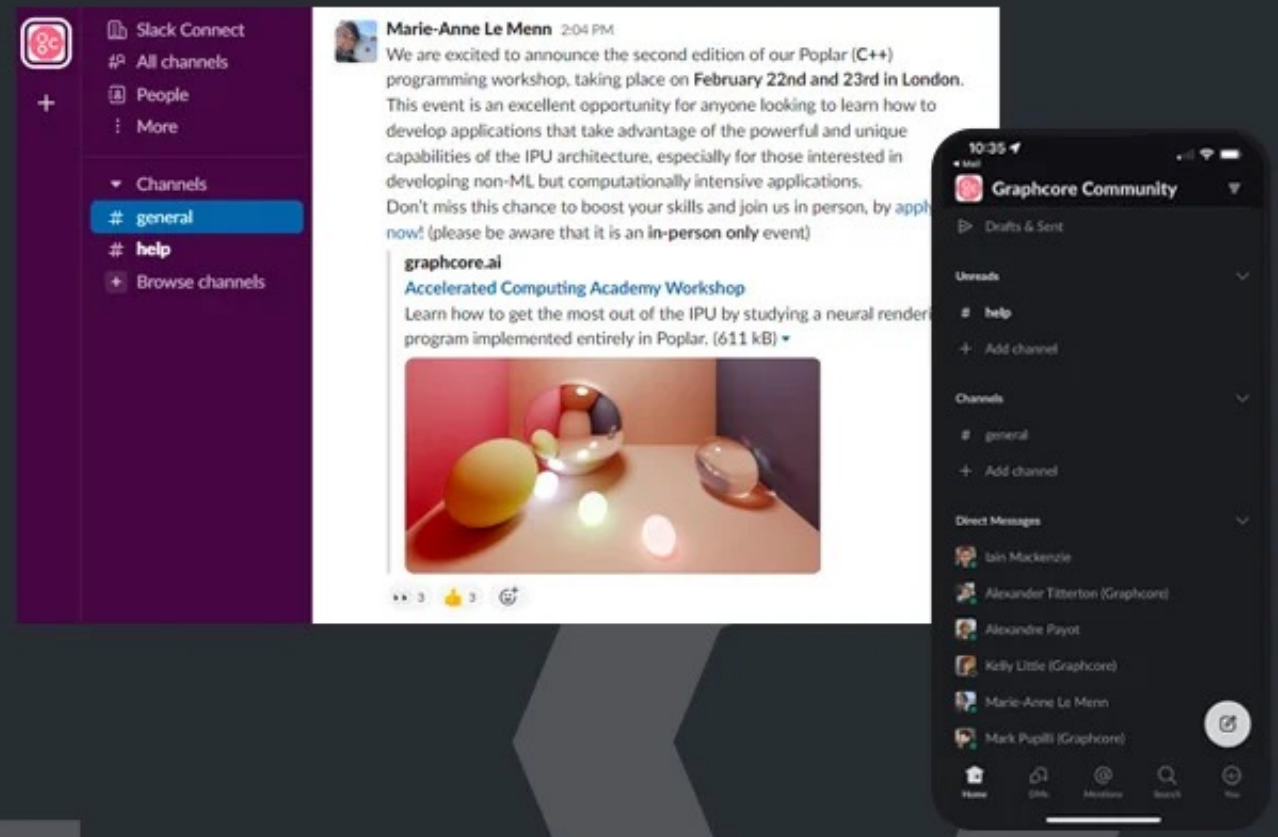
- ✓ Up to **$200,000** in IPU cloud credits
- ✓ Dedicated support from Graphcore's **GenAI experts**
- ✓ Bespoke training sessions and hands-on developer workshops
- ✓ Co-marketing promotion and awareness-building support
- ✓ Connect with startup founders, leaders, VCs and the ML community

APPLY NOW: *ipu.dev/startup*

GRAPHCORE

Join the conversation

slack



ipu.dev/join-community

ZACK ZWEIG

PRODUCT MANAGER
PAPERSPACE

DEMO: Deploying your Generative AI Apps in the Cloud

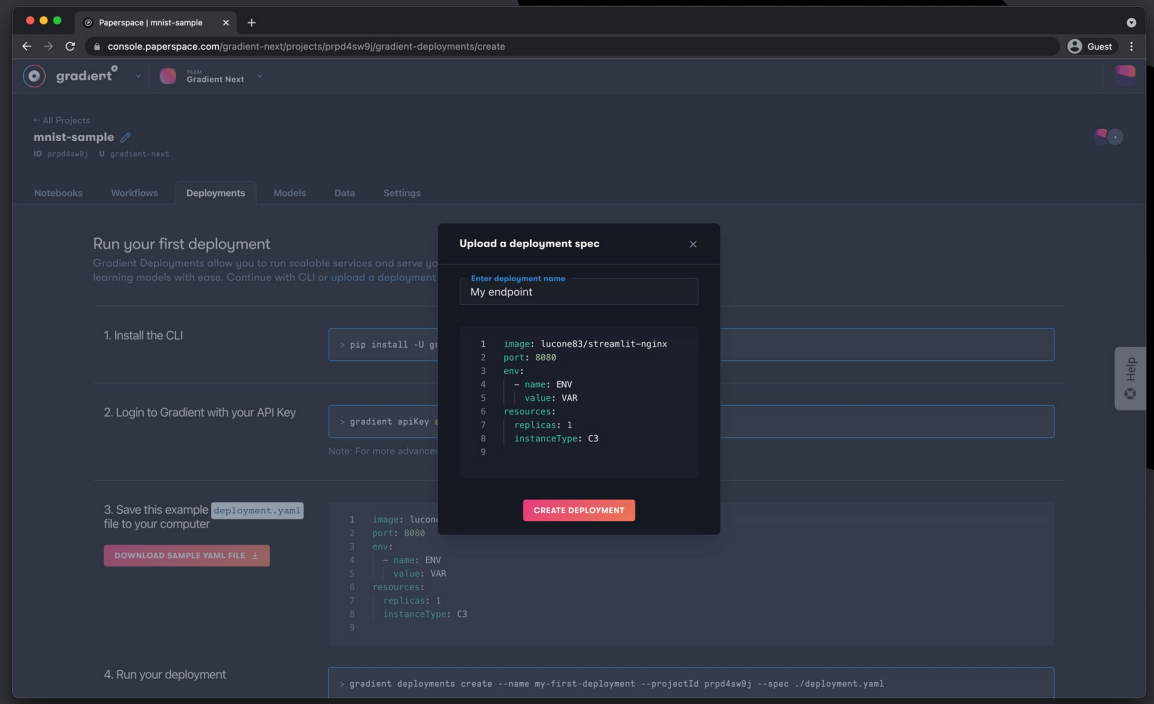
GRAPHCORE + *Paperspace*

COMING SOON

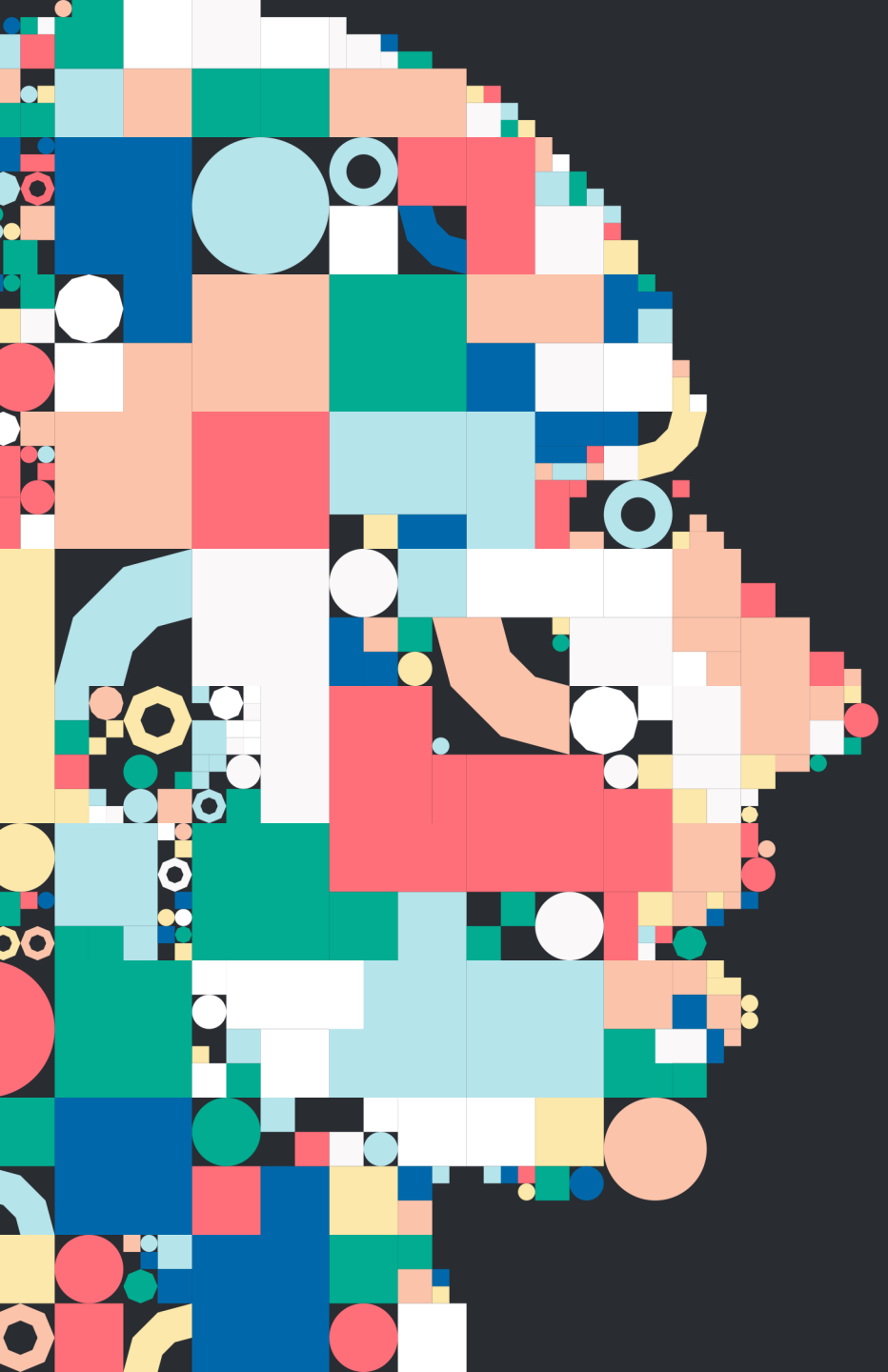
IPU DEPLOYMENTS

Deploy, manage and scale GenAI applications in the cloud with IPUs on Paperspace

Launching soon in Beta



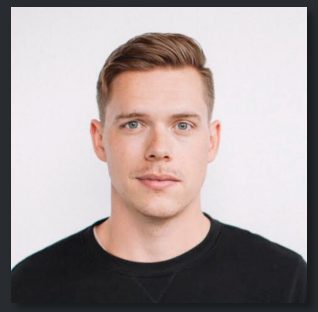
JOIN THE WAITING LIST *ipu.dev/deploy*



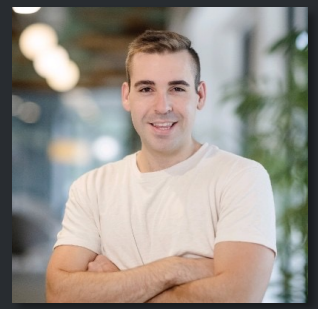
10 MIN BREAK

FIRESIDE CHAT / Q+A

Future of Generative AI: Challenges & Best Practices



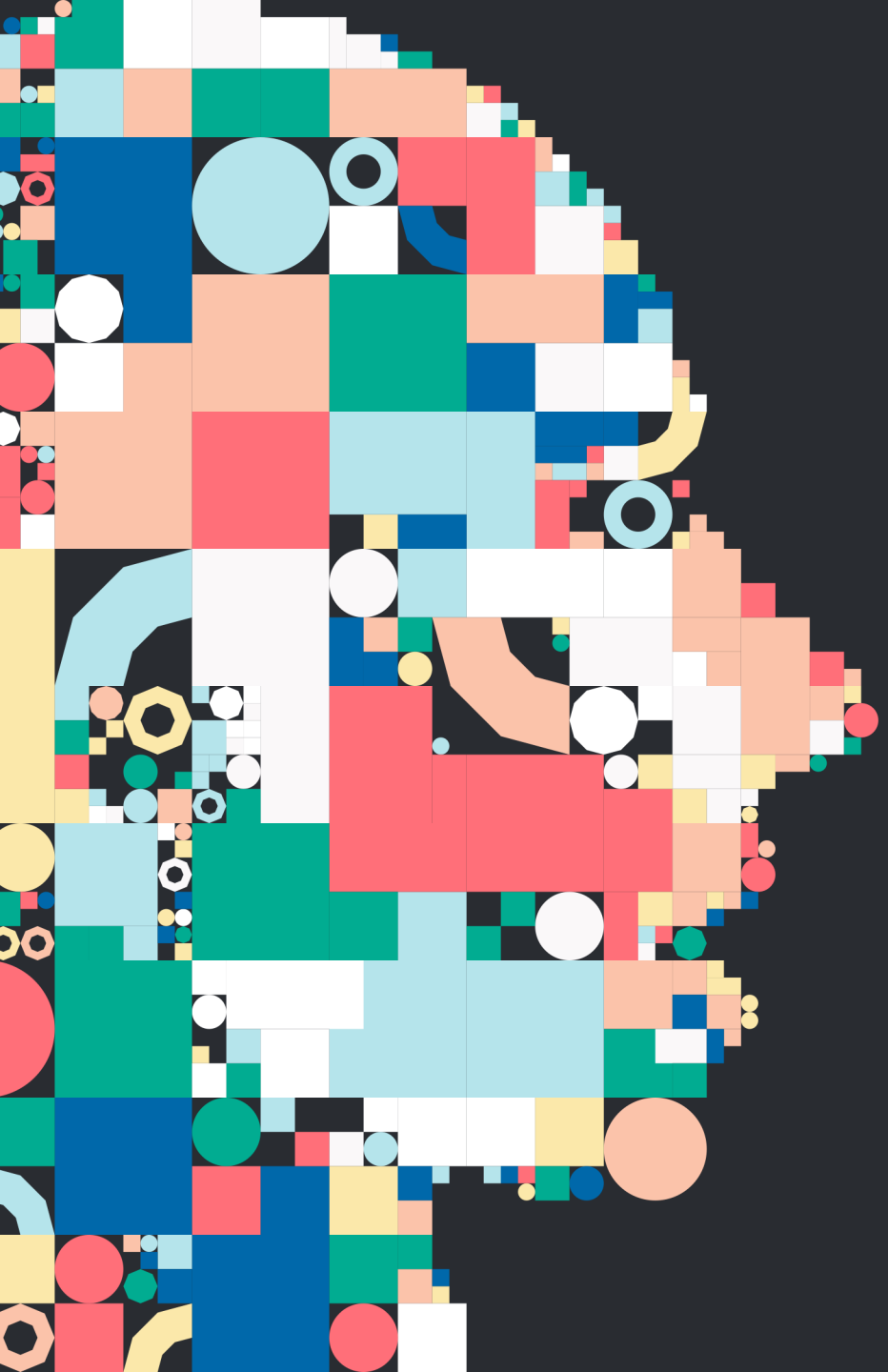
DILLON ERB
CEO
PAPERSPACE



ANDRIY MULYAR
FOUNDER & CTO
NOMIC AI

YULIYA SHCHERBACHOVA
DIRECTOR REVENUE ENABLEMENT
WEIGHTS & BIASES

HELEN BYRNE
VP SOLUTION ARCHITECTS
GRAPHCORE



THANK YOU!

JOIN US FOR NETWORKING DRINKS