Deploying Large Language Models with Ease

Hannes Hapke

Principal Machine Learning Engineer, Digits

Machine Learning
is at a crossroads…

Two issues

LLMs are great! But…

Do you really want to give your **precious data** to big corporations?

ML Projects are Icebergs

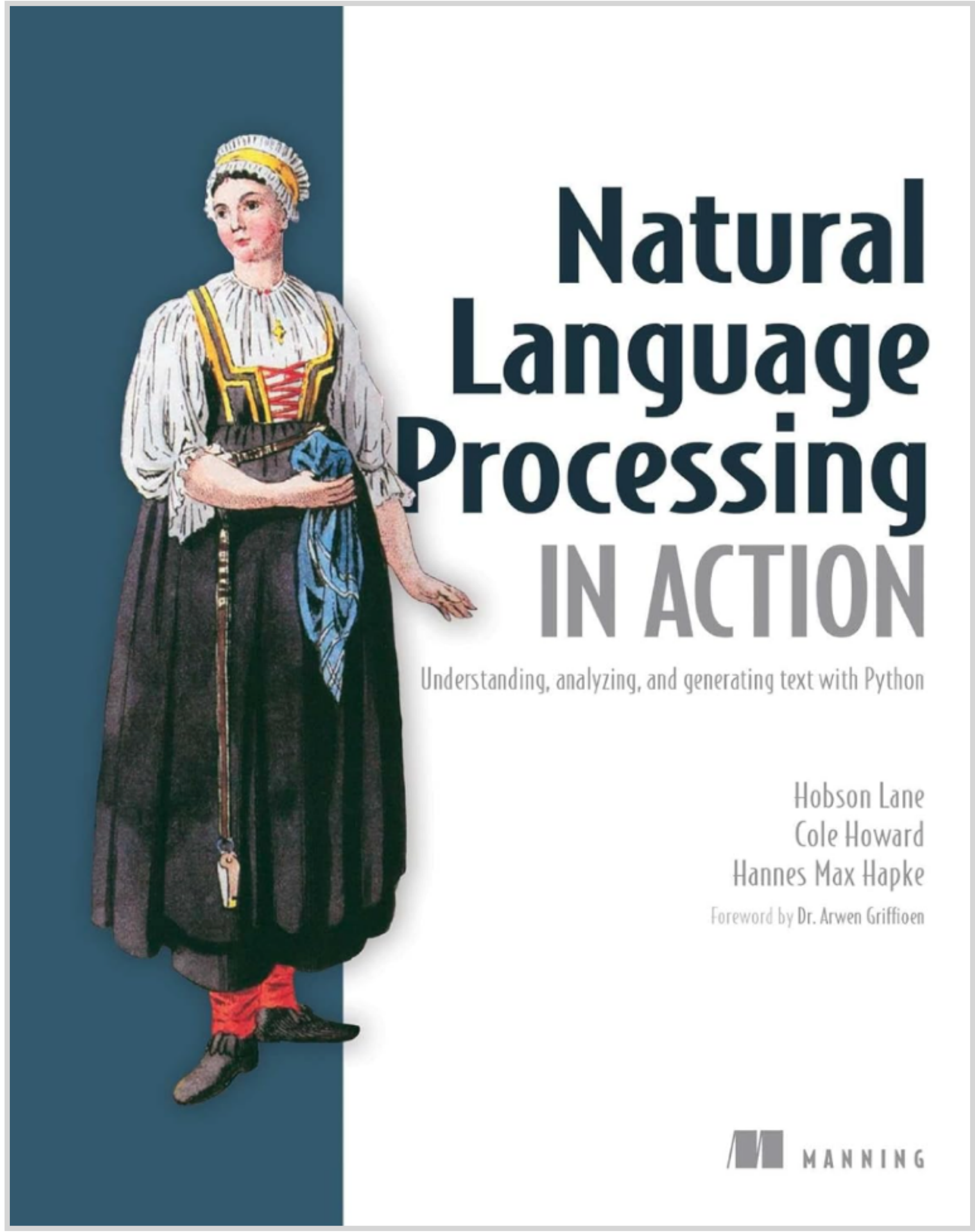

Visible to users
Focused on by bloggers
Important work to get models in production

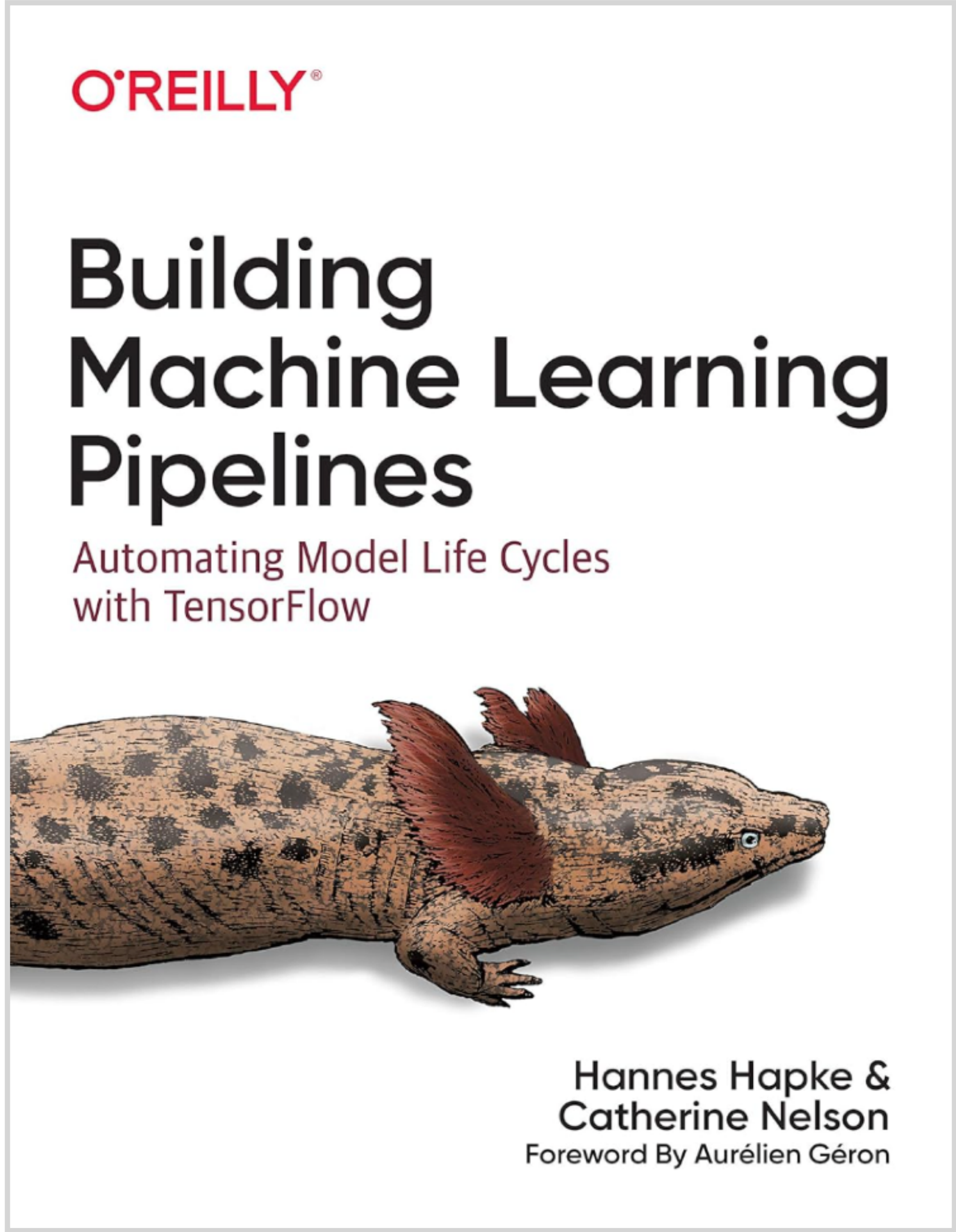

Visible to users
Focused on by
bloggers
Important work
to get models in
production
LLMs made the
iceberg deeper

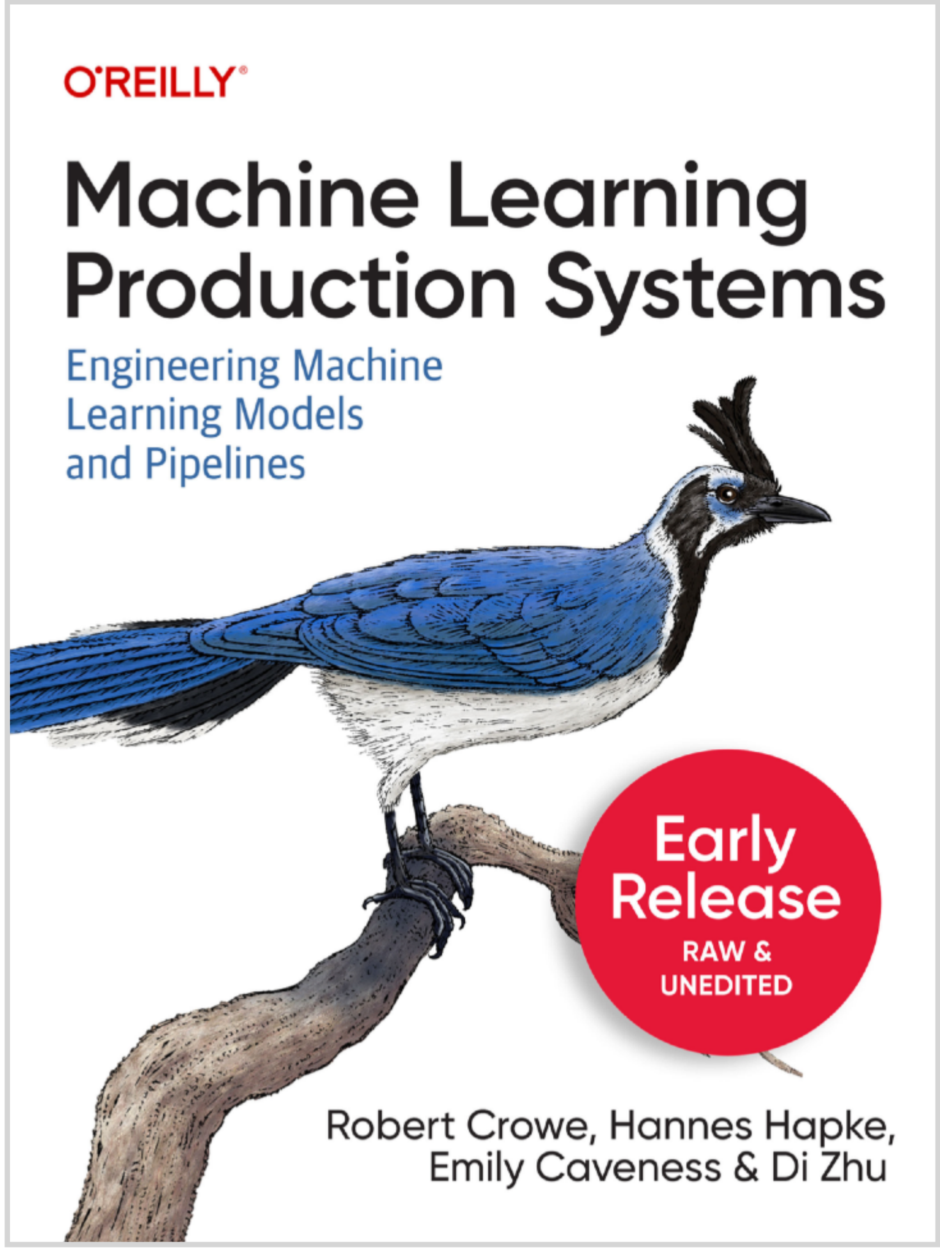
Icebergs can sink Titanics

Model APIs can shrink your iceberg into an ice cream cone

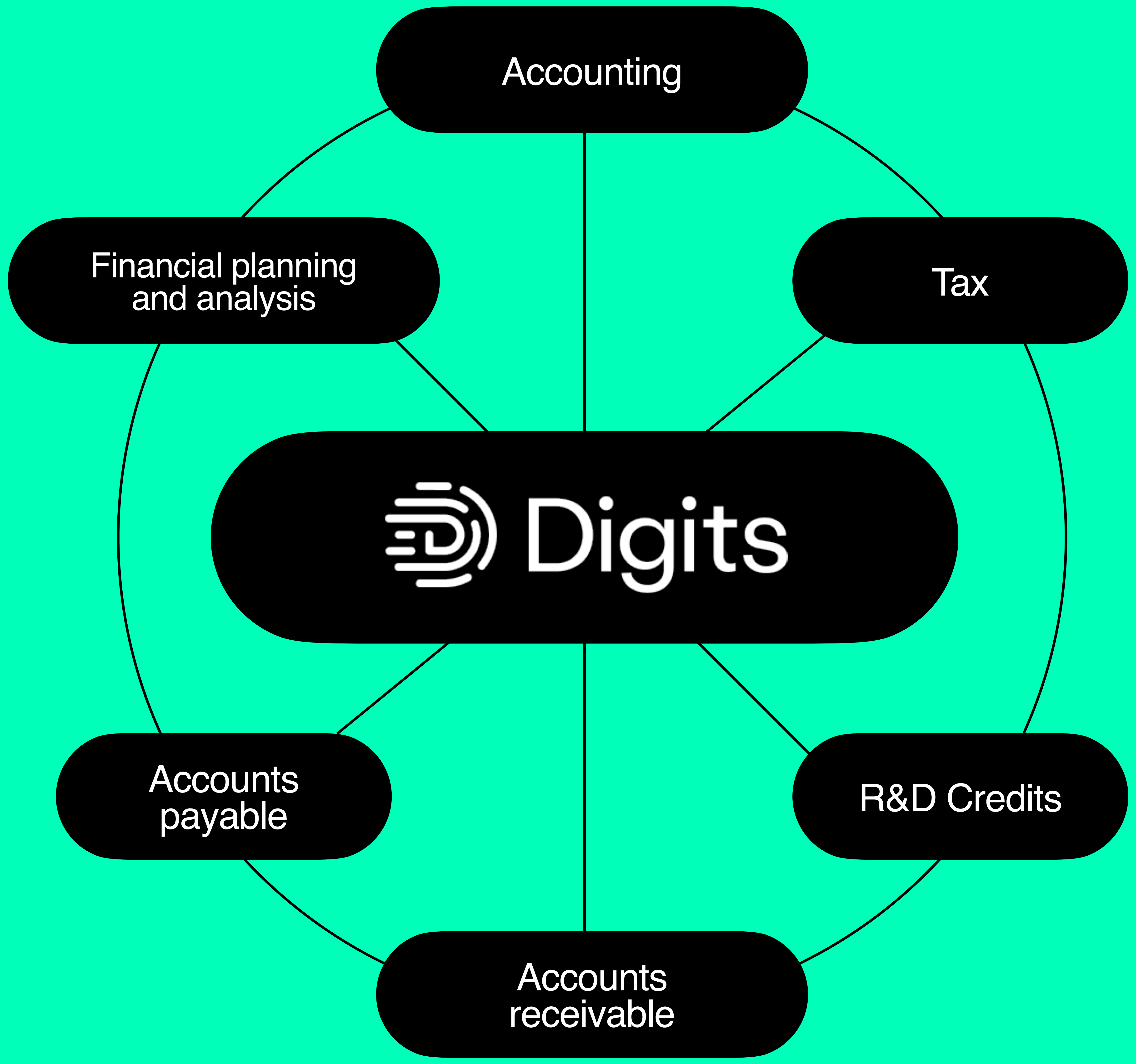
Hi, I'm Hannes.


Natural Language Processing
IN ACTION
Understanding, analyzing, and generating text with Python
Hobson Lane
Cole Howard
Hannes Max Hapke
Foreword by Dr. Arwen Griffioen
MANNING


O'REILLY
Building Machine Learning Pipelines
Automating Model Life Cycles with TensorFlow
Hannes Hapke &
Catherine Nelson
Foreword By Aurélien Géron


O'REILLY
Machine Learning Production Systems
Engineering Machine Learning Models and Pipelines
Early Release
RAW & UNEDITED
Robert Crowe, Hannes Hapke,
Emily Caveness & Di Zhu

Startup financial operations, powered by AI bookkeeping



How are we using LLMs?

Invoicing

New Invoice

Draft

Drop a sales contract, get an invoice!

New Invoice

Sent

XYZ Incorporated
Due June 30, 2024 $400.00

Total $400.00

Overdue

Nudge

Trade Connect
Due May 1, 2024 $400.00

8 Days Overdue 2

Signpost LLC.
Due May 30, 2024 $1,200.00

28 Days Overdue

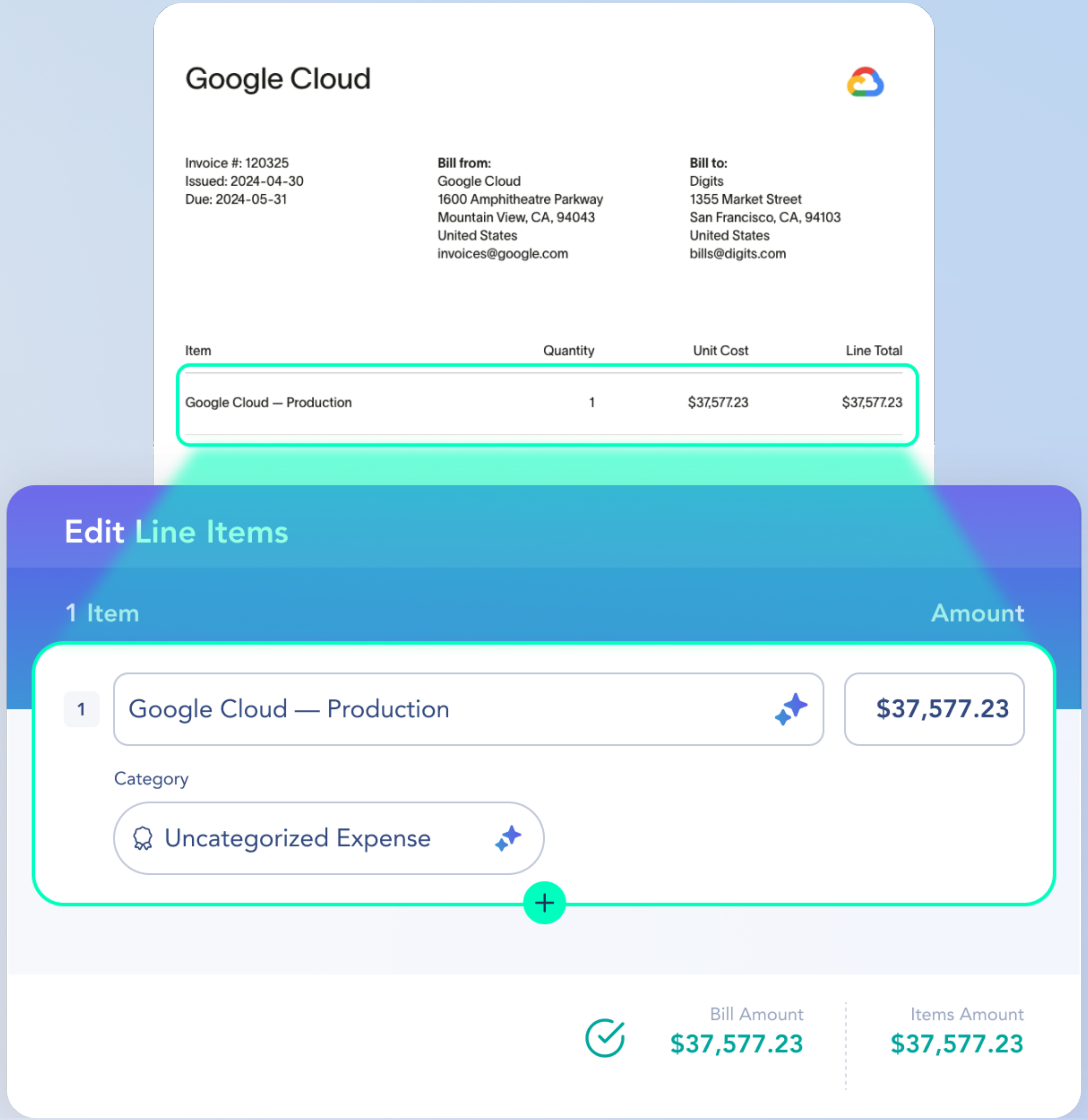
Total $1,600.00

Paid

Past 30 Days

Homebody Inc.
Due Mar 30, 2024 $800.00

Total $800.00



Understanding Documents

Vendor Details

Name	
Address	
Phone	
Website	
Social Media Links	
Wikipedia Site	
Description	
Short Description	
Keywords	
Is a National Brand	
Logo	

All runs in Production.

Why did the iceberg get bigger?

Resource intensive deployments

Constantly changing ML world

Difficult to fine-tune

Evaluation can be tricky

New tooling needed

Lessons Learned

Model Selection

Smaller is often better

Complex hosting 70b models

Latency can be a killer to LLMs

Very limited real-world benefits

MoE models seems limited


Focus on a specific model

Tooling and Infrastructure

GCP Vertex not designed for LLMs

Deployment is 10-100x expensive

CUDA

Choose your hosting framework wisely

NVIDIA, Triton, TitanML, vLLM

Inference Optimizations

Parallelization

Efficient Attention

Quantization

Continuous Batching

Parallelization

- Multi-GPU / multi-node inference
- Quick explosion of costs

Parallelization

Costs increase: 2-5x

Efficient Attention

KV Caching is King/Queen

No recomputation of previous tokens needed

E.g. PagedAttention

Example: 70b Model

$\text{Tokens} \times n_{layers} \times n_{kv\_heads} \times d_{heads} \times \text{sizeof(dtype)}$

Weights: ~130 GiB

KV Cache: ~160 GiB

Quantization

Reduce memory requirements, reduce weights + KV cache

Generalized Post-Training Quantization (GPTQ)

- Balanced between compression gains and inference speed
- Focuses on GPU inference and flexibility in quantization levels

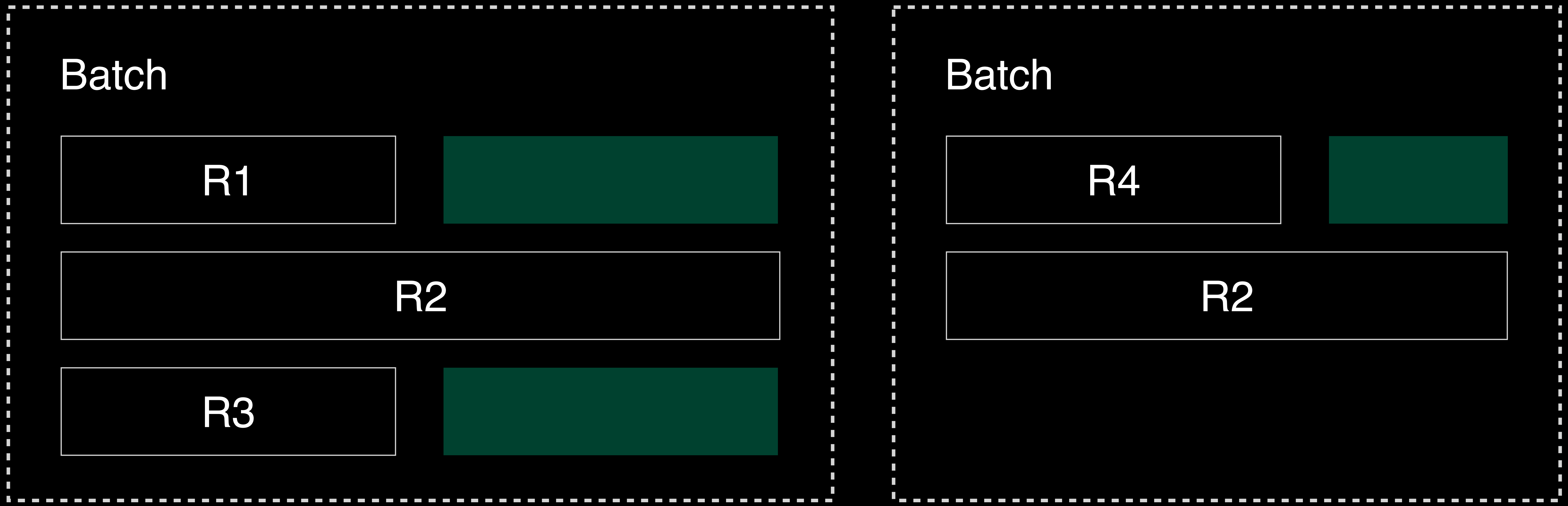
Activation-Aware Weight Quantization (AWQ)

Observes activations for weight quantization

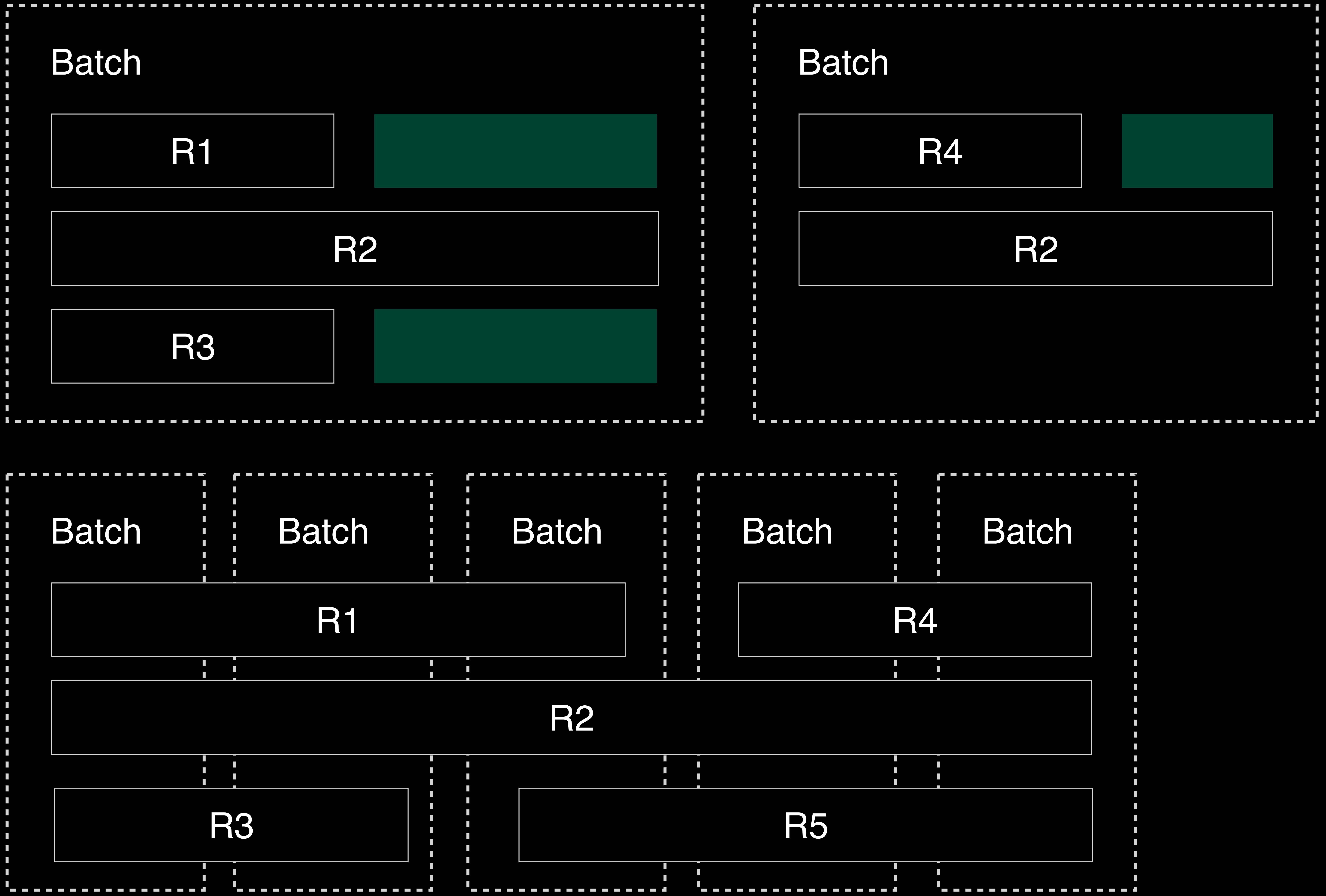
High quantization performance for instruction-tuned LLMs

Quantitization

Costs reduced: 90%
Very project specific



Batching



Continuous Batching

Batching

Improvements: 2x

Soon: 3-4x

Deployment Strategies

Batching

Easier to host very large LLMs

Less cost intensive

Streaming

- Idle time will be expensive
- More cost intensive
- Latency matters
- Users love the immediate results

Other Lessons Learned

Lessons Learned

- Users got used to streaming
- Serving Streaming Mode is another beast
- Latency to first token is less important than overall task completion

Conclusions

Smaller is often better

Consider the model latency

Batching beats Streaming

Don’t follow every trend, focus

OS LLMs are a long game

Thank you!

digits.com