Intel® Gaudi® in Action: Solving Real-World Challenges with Fine-Tuned Language Models

Burak Aksar - Spiky.ai/Founder

Eduardo Alvarez - Intel/Senior AI Solutions Engineer

oneAPI

SPIKY.AI

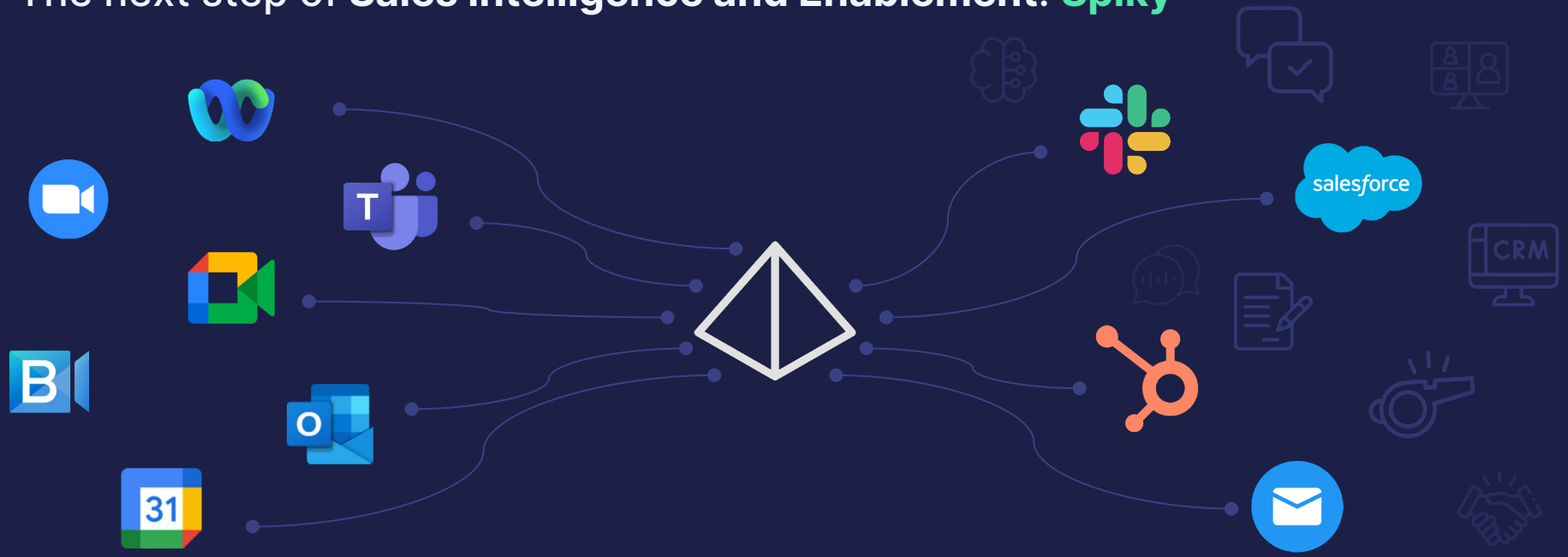
AI Supercharged Sales

spiky.ai

info@spiky.ai

company/spiky-ai spikyai spikyai

The next step of **Sales Intelligence and Enablement: Spiky**



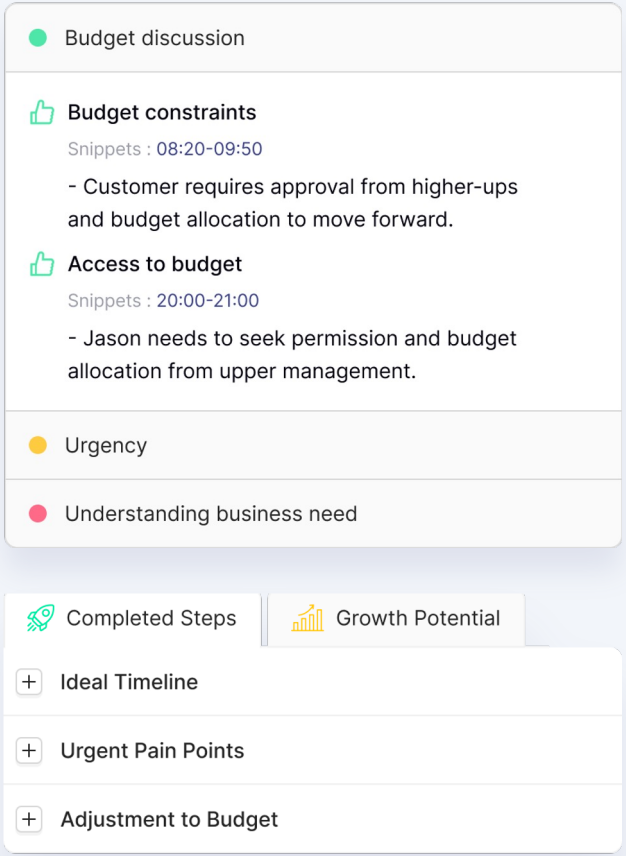
Calendar, CRM, and video conferencing tool integrations end-to-end automation

Automatically generated actionable sales feedback

Valuable people insights with moment detection

Summarization and followup generation

Coaching for every stage of sales



Team, Deal, Pipeline Visibility

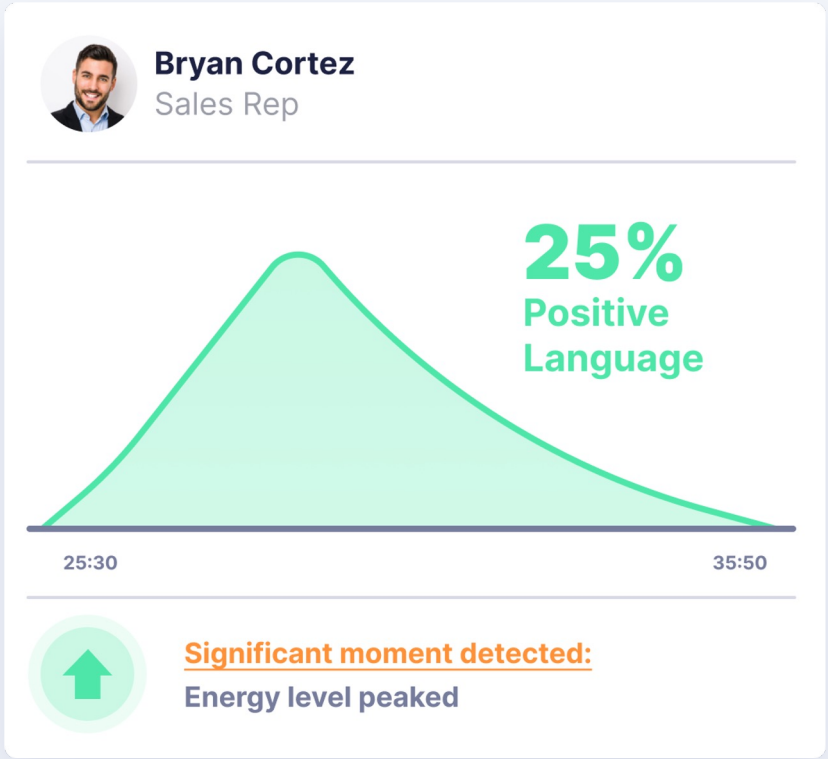
Real-time Playbook

Auto insights for each sales stage

Utilize popular sales frameworks or build your own

Moments and EQ Intelligence

Next-Gen Meeting Automation



Team, Deal, Pipeline Visibility

Sales-stage Aware

Moments and EQ Intelligence

90% of communication is non-verbal

Understand customer moments

Next-Gen Meeting Automation

Team, Deal, Pipeline Visibility

Sales-stage Aware

Moments and EQ Intelligence

Ready to go in 10 seconds

Organize and gain insights from meetings in one place

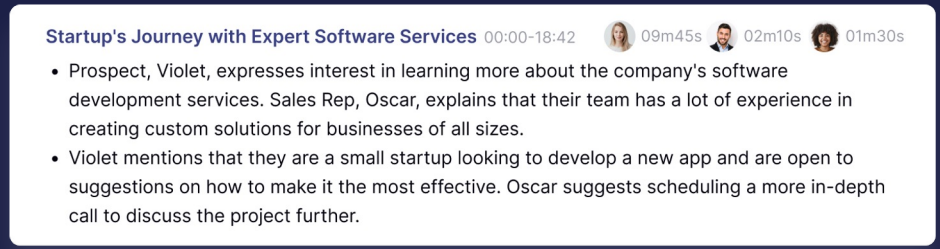
Spiky is platform agnostic

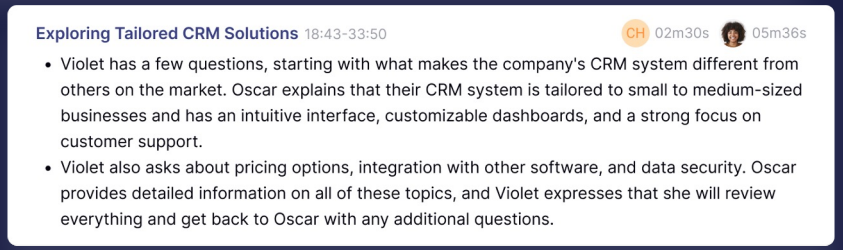
Personalizing Revenue Teams' Communication with Spiky

Spiky leverages more than 25 models targeting vocal, language, and visual metrics

Some internal datasets comprising:

- Sales call transcripts
- Industry-specific jargon
- Customer feedback
- Successful email correspondences





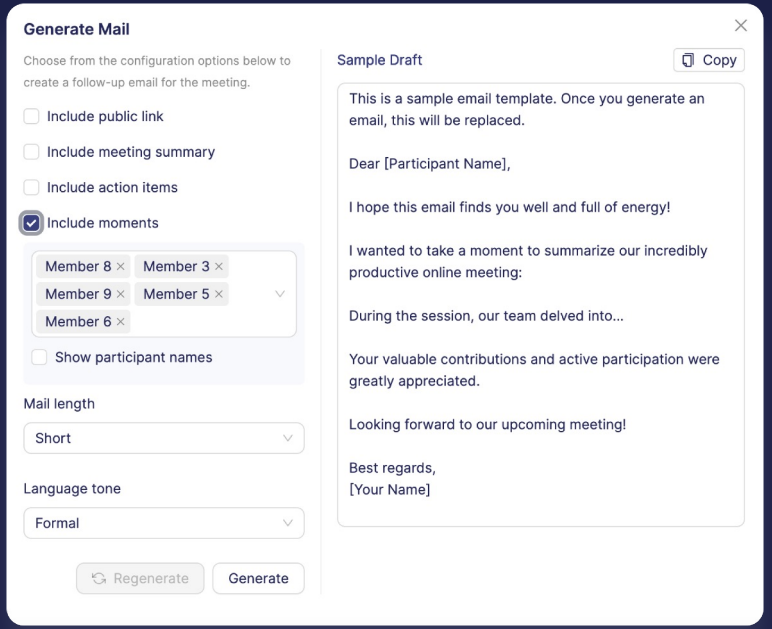
Personalizing Revenue Teams' Communication with Spiky

Spiky leverages more than 25 models targeting vocal, language, and visual metrics

Developing a model adept in sales contexts:

- Generating precise and relevant follow-up emails
- Explaining customer behavior

  for certain metric changes
- Enhanced clarity; tailored to customers' style
- Significant increase in sales conversions



Enhancing Sales Communication with AI: The Spiky Initiative

Current approach:

Create "Spiky Moments Dataset":

- Adapt from existing open-source spoken dialogue datasets
- Integrate customer ratings for enhanced relevance

Fine-tune open-source LLMs with **Spiky Moments Dataset** using **Gaudi V1 and Nvidia A10**

- Intel/neural-chat-7b-v3-1 Nvidia A10
- mistralai/Mistral-7B-Instruct-v0.1 Nvidia A10
- meta-llama/Llama-2-7b-chat-hf Gaudi V1

gpt-3.5-turbo

In this part of the meeting transcript, Danilo and Sameed discuss the launch setting for a campaign. They consider allowing users to select their preferred locale or skip such users for the campaign. Sameed mentions adding a dropdown menu in the preview section to select the desired locale, and they discuss the possibility of adding more details or creating a new type for localization or multi-language experiments....

llama-7b–spiky-moment

Danilo and Sameed discuss a project that involves localized campaign analytics. Their focus is on creating an accurate campaign based on specific rules, user-defined rulesets, and exclusions. They touch upon previous work and how to best approach the analytics page. They plan to keep the page simple for now, updating and adding features based on future requirements and feedback, as well as managing expectations with the client....

- Improved task performance with domain specific fine-tuning
- Achieved faster training using Gaudi V1 (compared with Nvidia A10)
- Cheaper inference using 4-bit quantization against other commercial APIs

Next Steps

- **Impact of different dataset sizes to** the model performance

- **Hardware-level optimization** for the model training time

- Explore other **further optimizations** for the model performance

- **Explore Gaudi V2.0**

Try today!

SPIKY.AI **Supercharge Sales Performance with AI**

AMROCK VENTURES techstars_ TechOne m]x[v DormRoomFund ERA GLOBAL aws activate NVIDIA INCEPTION PROGRAM Microsoft for Startups

intel Ai summit

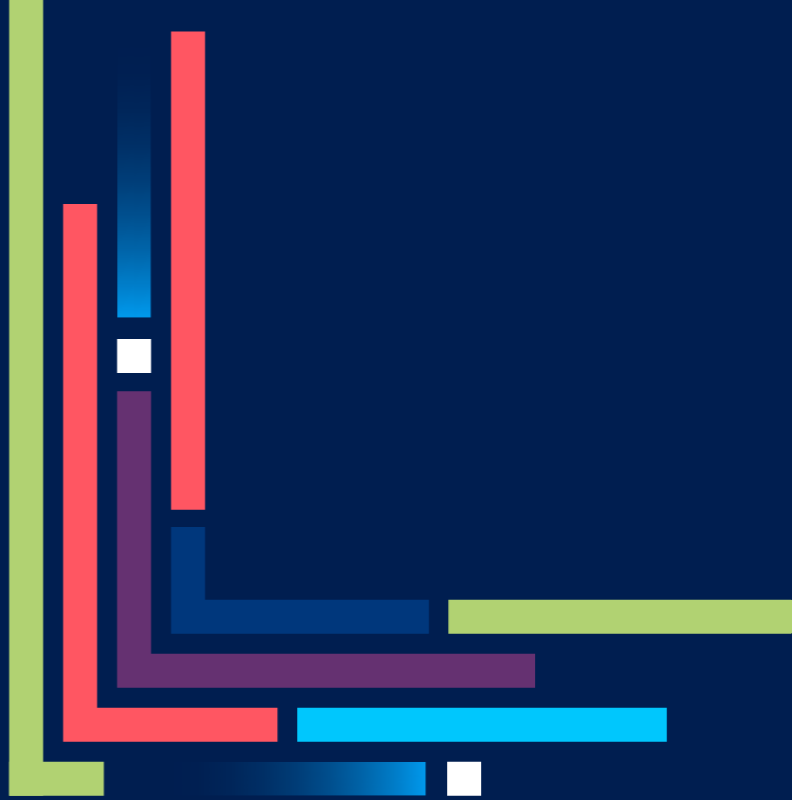
Bringing AI Everywhere

Opportunities on Gaudi for Developers

Intel

Eduardo Alvarez

Senior AI Solutions Engineer



Spanning the Stack, Across the Lifecycle

Engineer Data → Create Machine Learning & Deep Learning Models → Deploy

Container Repository	oneAPI powered	MLOps	Developer Sandbox	Annotation/Training/Optimization
oneContainer	**AI Reference Kits**	**Cnvrg.io**	**Intel® Developer Cloud**	**Intel® Geti™**

Connect AI to Big Data — Apache Spark — BigDL (previously "Analytics Zoo")

Data Analytics Scale

MODIN · SciPy · pandas · NumPy

Optimized Frameworks and Middleware

TensorFlow · PyTorch · mxnet · PaddlePaddle · scikit-learn · ONNX · LightGBM · XGBoost · CatBoost

Optimize Models

Automate Model Tuning AutoML

Automate Low-Precision Optimization

SigOpt

Intel Neural Compressor

OpenVINO™ Toolkit

Write Once

Deploy Auto-Optimized

Anywhere

SYCLomatic · oneDAL · oneDNN · oneCCL · oneMKL · SynapseAI™

intel ATOM · intel CORE · intel XEON · intel ARC GRAPHICS · intel DATA CENTER GPU · intel GAUDI

intel AI summit

Note: not all components are necessarily compatible with all other components in other layers

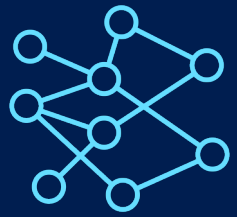
intel®
Developer Cloud

intel Ai summit

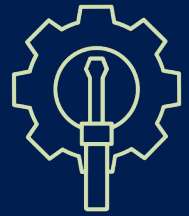
Making the SoTA Accessible

Focus on Efficient SoTA with Gaudi2

LoRA has been recently enabled on Gaudi2 through easy-to-use Hugging Face APIs



Reduces the complexity of models, making them easier to train, with little to no degradation in performance.



Fine-tuning with LoRA on Gaudi2 leads to more efficient memory and compute efficiency.

AI Libraries	Hugging Face, DeepSpeed
HPU API	SynapseAI®
Platform	intel Developer Cloud
Hardware	intel Xeon, intel Gaudi

intel AI summit

E2E Llama 2 on Gaudi with Hugging Face

Llama2-7B

+

Dataset: openassistant-guanaco
9.85K Rows

+

Optimum Habana

PEFT (LoRA)

deepspeed

+

$10.42 / hour

intel Developer Cloud

intel GAUDI

8 Gaudi2® HL-225H mezzanine cards with 3rd Gen Xeon® processors, 1 TB RAM, 30 TB disk

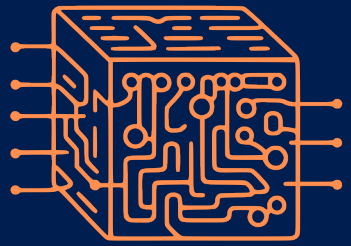
=

Tuned Llama in ~5min for ~$1

intel AI summit

GPU to HPU Model Portability

GPU to HPU Model Portability



The ultimate way to escape vendor lock and enable AI engineering fluidity across powerful accelerators

HPU API	Habana GPU Migration APIs
Platform	intel Developer Cloud
Hardware	intel Xeon, intel Gaudi

intel AI summit

What's next for you?

Check out the Intel Developer Cloud!

visit
cloud.intel.com

intel Ai summit