

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

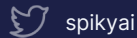
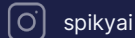
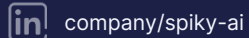
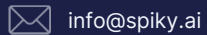
Burak Aksar - [Spiky.ai](https://spiky.ai)/Founder

Eduardo Alvarez - Intel/Senior AI Solutions Engineer

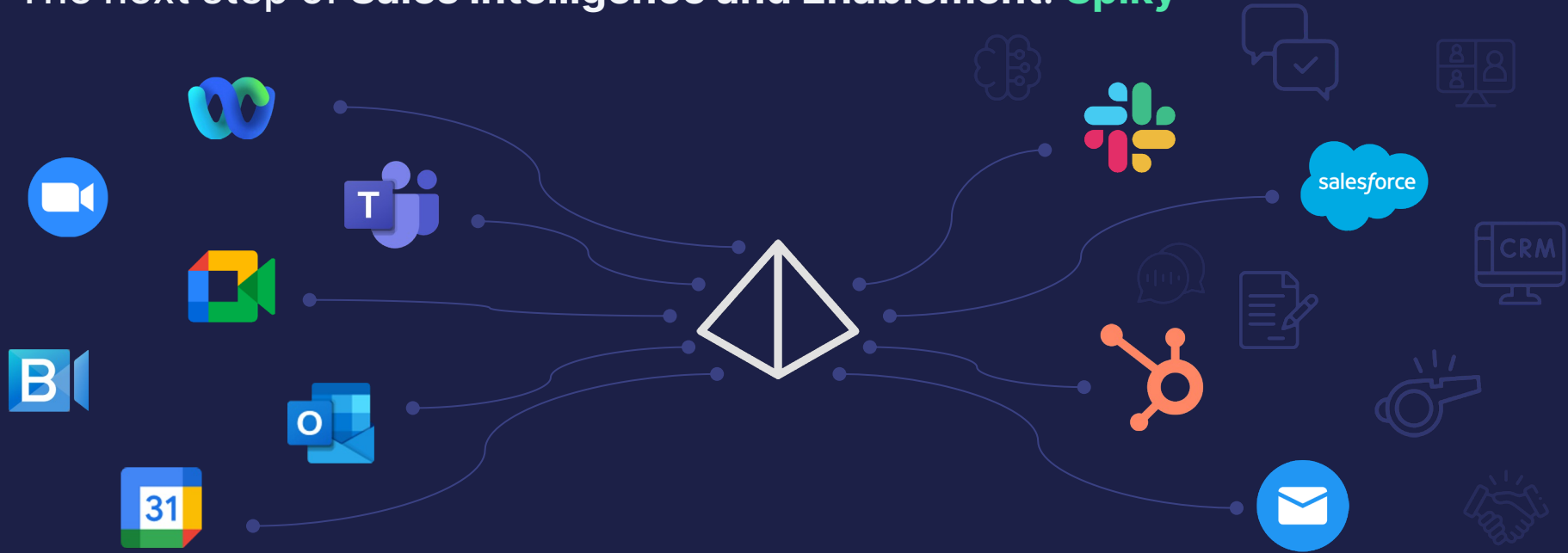




AI Supercharged Sales



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


 Budget discussion


 **Budget constraints**
Snippets : 08:20-09:50



- Customer requires approval from higher-ups and budget allocation to move forward.


 **Access to budget**
Snippets : 20:00-21:00


- Jason needs to seek permission and budget allocation from upper management.


 Urgency

 Understanding business need

 Completed Steps  Growth Potential

 Ideal Timeline

 Urgent Pain Points

 Adjustment to Budget

Team, Deal, Pipeline Visibility

Real-time Playbook

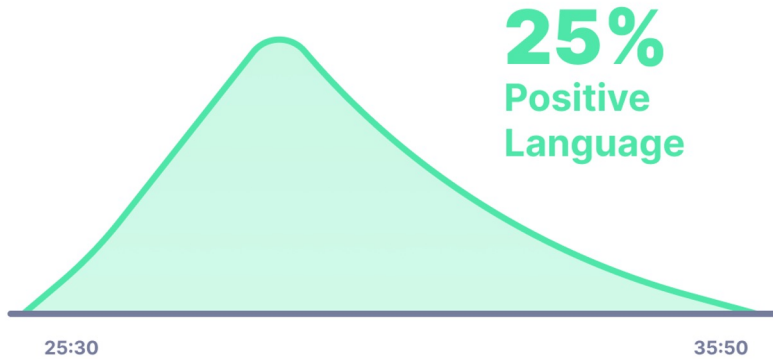
Auto insights for each sales stage
Utilize popular sales frameworks or
build your own

Moments and EQ Intelligence

Next-Gen Meeting Automation



Bryan Cortez
Sales Rep



Significant moment detected:
Energy level peaked

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

Startup's Journey with Expert Software Services 00:00-18:42  09m45s  02m10s  01m30s

- Prospect, Violet, expresses interest in learning more about the company's software development services. Sales Rep, Oscar, explains that their team has a lot of experience in creating custom solutions for businesses of all sizes.
- Violet mentions that they are a small startup looking to develop a new app and are open to suggestions on how to make it the most effective. Oscar suggests scheduling a more in-depth call to discuss the project further.

Exploring Tailored CRM Solutions 18:43-33:50  02m30s  05m36s

- Violet has a few questions, starting with what makes the company's CRM system different from others on the market. Oscar explains that their CRM system is tailored to small to medium-sized businesses and has an intuitive interface, customizable dashboards, and a strong focus on customer support.
- Violet also asks about pricing options, integration with other software, and data security. Oscar provides detailed information on all of these topics, and Violet expresses that she will review everything and get back to Oscar with any additional questions.

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

Generate Mail

Choose from the configuration options below to create a follow-up email for the meeting.

- Include public link
- Include meeting summary
- Include action items
- Include moments

Member 8 × Member 3 ×
Member 9 × Member 5 ×
Member 6 ×

Show participant names

Mail length
Short

Language tone
Formal

Regenerate Generate

Sample Draft

Copy

This is a sample email template. Once you generate an email, this will be replaced.

Dear [Participant Name],

I hope this email finds you well and full of energy!

I wanted to take a moment to summarize our incredibly productive online meeting:

During the session, our team delved into...

Your valuable contributions and active participation were greatly appreciated.

Looking forward to our upcoming meeting!

Best regards,
[Your Name]

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!



Supercharge Sales Performance with AI

AMROCK
VENTURES

techstars_

TechOne



DormRoomFund



aws activate

nVIDIA
INCEPTION PROGRAM



intel[®]Ai
summit

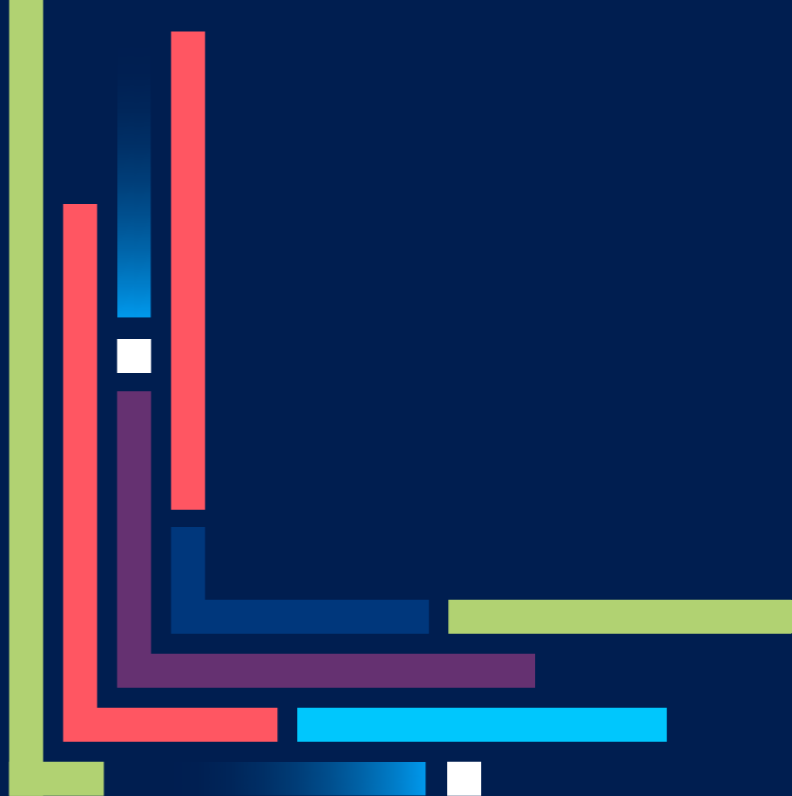
Bringing AI Everywhere

Opportunities on Gaudi for Developers

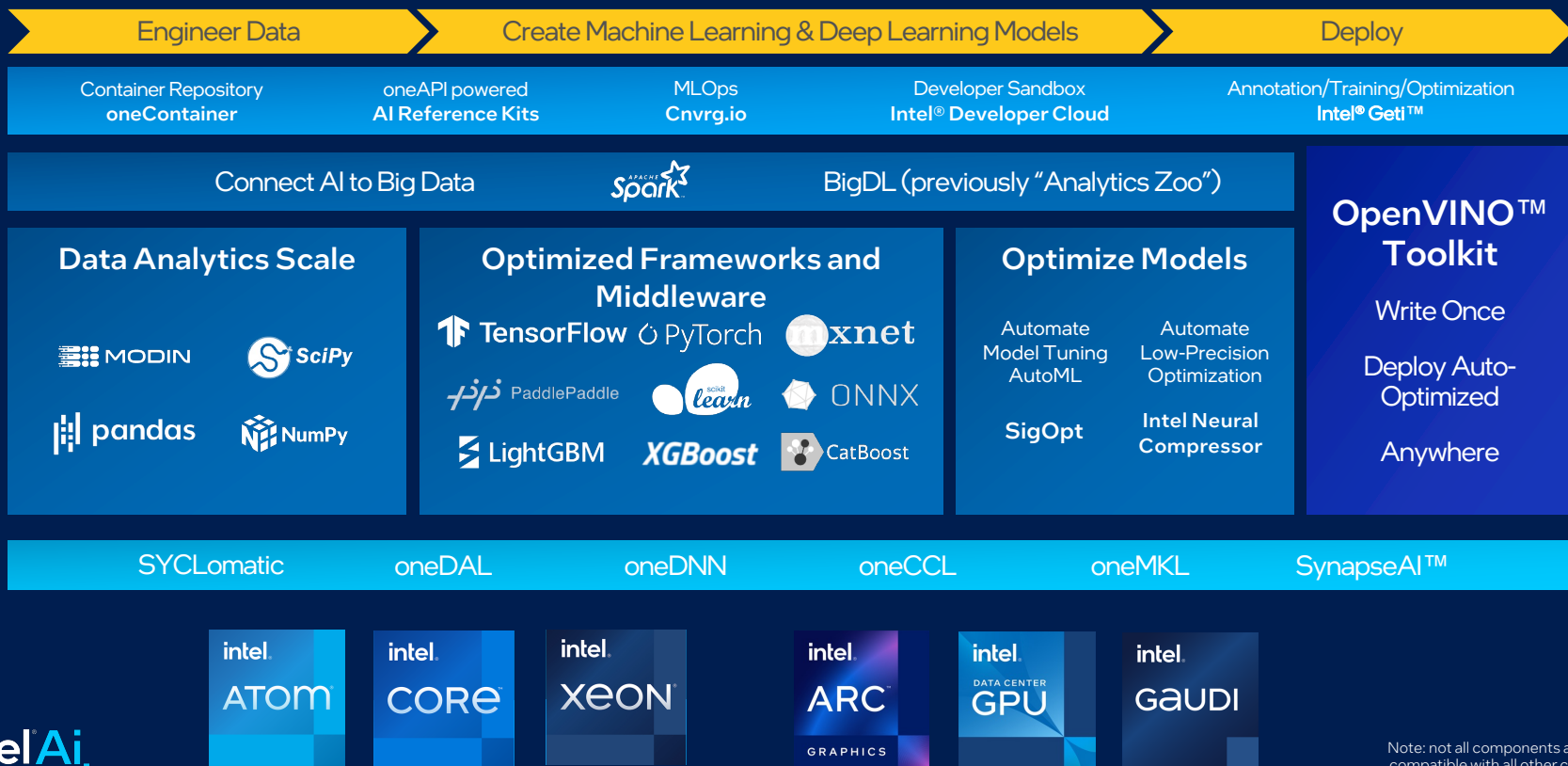
Intel

Eduardo Alvarez

Senior AI Solutions Engineer



Spanning the Stack, Across the Lifecycle



intel[®] Developer Cloud

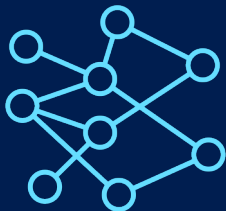


Making the SoTA Accessible

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 Ai
summit



E2E Llama 2 on Gaudi with Hugging Face



Llama2-7B



Dataset:
openassistant-
guanaco
9.85K Rows

A yellow emoji with its hands clasped, representing Optimum Habana.

Optimum Habana

A yellow emoji with its hands clasped, representing PEFT (LoRA).

PEFT (LoRA)

A network diagram icon representing DeepSpeed.

deepspeed



\$10.42 / hour

intel
Developer Cloud

The Intel Gaudi logo, consisting of the word 'intel' in white and 'GAUDI' in white on a dark blue background.

intel
GAUDI

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

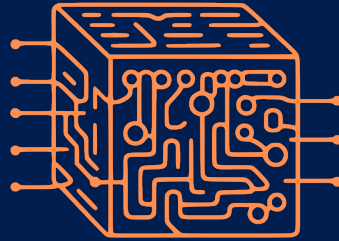


Tuned
Llama in
~5min for ~\$1



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



What's next for you?



Check out the Intel
Developer Cloud!

visit
cloud.intel.com