Intel -Denvr Habana Gaudi Customer Questionnaire

(Consult Intel AI Solutions Team if answers are unclear or qualification is uncertain.)

Document version 1.0

Date: June 26th, 2024

# Intel - Denvr sales contact points:

|  |  |
| --- | --- |
| Intel - Denvr Account Executive | Juan Demaestri [juan.demaestri@intel.com](mailto:juan.demaestri@intel.com) |
| Intel - Denvr AI Advisor | Sumesh Sukumar [sumesh.sukumar@intel.com](mailto:sumesh.sukumar@intel.com) |
| Intel AI Solutions Team | [smg.americas.ai.team@intel.com](mailto:smg.americas.ai.team@intel.com) |
| Intel AI Business Developer | Zahler, Deborah “Deb” [deborah.zahler@intel.com](mailto:deborah.zahler@intel.com) |
| Intel AI Technical Sales Manager | Pandit, Milind S [milind.s.pandit@intel.com](mailto:milind.s.pandit@intel.com) |

# Customer evaluation questionnaire

|  |  |
| --- | --- |
| Question | Answer |
| What is the main AI use case, e.g.:   * (large) language models * generative AI * computer vision | Large language model |
| What are the key relevant models, e.g.:   * Llama2-70b * ResNeXt-101   Include model sizes and GitHub links, if applicable.  Which models are relevant for the evaluation? | Llama3 8/70b |
| For the specific model, do you want to use it for:   * Inference * Fine-tuning * Training   Include relevant info about dataset, if applicable. | Training and Inference |
| If the specific model you want to train is not publicly available, are you willing to share it under CNDA? | No |
| If not, is there a publicly available proxy we should refer to? | NA |
| Does the model rely on dynamic shapes? | No |
| Does your model use custom CUDA operators, and if so, what functions do they perform? | No |
| Are you okay with using Megatron DeepSpeed for training LLMs at scale? | <Input required> |
| Do you have dependencies on other software libraries, deployment frameworks, e.g.:   * FineTuning with LoRA, FSDP, … * Deployment with Triton, TGI, VLLM, … | *<Input required>* |
|  |  |
| Do you know how much it costs you to train or do inference with your model? Has compute become a pain point for you? | NA |
| What is the size of your existing deep learning model infrastructure? This refers to hardware resources currently in use (generation and number of GPUs, storage requirements, etc…) | NA |
| Can you communicate the planned or required growth of the AI infrastructure to satisfy your existing or near-future use cases? | Yes, we will inform in near future for the use cases |
| Can you explain timeline for decision making process (POC completion, RFP, deployment, etc…) | Will take 1 week to test the model |
| What performance/cost benefits would justify changing a few lines of code to your training or inference script? | Performance Benefit: Reducing the time spent on data loading and preprocessing can speed up the training and inference processes.  Cost Benefit: Faster data handling reduces the overall runtime, leading to lower computational costs. |
| What are the success criteria for the evaluation criteria that, if met, will result in purchase? E.g.   * Performance / cost competitiveness * Specific performance level (absolute or in relation to competitive products) * Compatibility with required models * Ease of use | Success criteria will be measure by performance and the required inference we received from the model with greater accuracy. |

# Appendix – supported models reference

These are example models we recommend customers to start with.

It is not a closed list, just a few examples in each category.

Model repositories and references:

* [HuggingFace optimum Habana](https://huggingface.co/docs/optimum/habana/index)
  + [Gaudi-Validated HuggingFace models](https://huggingface.co/models?other=or:bert,stable-diffusion,albert,wav2vec2,vit,swin,t5,distilbert,roberta,gpt2,clip,bloom,gpt_bigcode,opt,esm,gptj,gpt_neox,bridgetower,llama,peft,llama-2,codegen,falcon,mpt,bart,mistral,blip,mixtral,whisper,phi,speecht5,persimmon,gemma,qwen2,owlvit,protst,audio-spectrogram-transformer)
* [Habana Model Catalog](https://developer.habana.ai/habana-catalog/)
* [Models Performance](https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html)

## Large Language Models

|  |  |  |
| --- | --- | --- |
| **Inference** | **Fine-Tuning** | **Full model pre-training** |
| Llama3 8/70b | Llama3 8/70b | Llama3 8/70b |
| Llama2 7/13/70B | Llama2 7/13/70B | Llama2 7/13/70B |
| Mistral 7B, Mixtral 8x7B | Mistral 7B, Mixtral 8x7B | Mixtral 8x7B |
| Falcon-40/180B | Falcon-40/180B | GPT2 |
| GPT-J 6B | GPT-J 6B |  |
| StarCoder | GPT-NeoX |  |
| Gemma, Phi, MPT |  |  |

## Generative AI and multi-modal models (beyond LLMs)

|  |  |  |  |
| --- | --- | --- | --- |
| **Domain** | **Inference** | **Fine-Tuning** | **Full model pre-training** |
| Speech to text | Wav2Vec, Whisper | Wav2Vec, Whisper |  |
| Text to Speech | SpeechT5 |  |  |
| Image to text | Llava, Llava-next, Blip |  |  |
| Image-text training | Clip, BridgeTower | Clip, BridgeTower |  |
| Text to image generation | Stable Diffusion 2.1/XL | Stable Diffusion 2.1/XL | Stable Diffusion 2.1 |
| Protein folding | ESMFold |  |  |

## Computer Vision

|  |  |
| --- | --- |
| **Inference** | **Full model training** |
| ViT, Swin (Visual Transformers) | ViT |
| ResNet, RexNeXt | ResNet, RexNeXt |
| YoloX | YoloX |
| Unet2D, Unet3D | Unet2D, Unet3D |
| SSD | SSD |