Seekr: Building Trustworthy LLMs for Evaluating & Generating Content at Scale

Intel® Developer Cloud enables AI companies to build, train, & deploy innovative solutions on a leading-edge, cost-effective AI acceleration platform.

intel. Developer Cloud

A cloud development environment with access to the latest Intel hardware & software to build & test AI & HPC applications for cloud, enterprise, client & edge deployments.
cloud.intel.com

Digital content drives marketing strategies and advertising revenue, but ensuring brand safety and suitability is challenging.

Building a trustworthy brand requires a thoughtful content strategy. Content production and media alignment must be brand-safe and suitable for target audiences. Yet, the tremendous volume of digital content and wide variety of channels makes it nearly impossible for companies to ensure that what they distribute consistently upholds brand standards.

Seekr is helping solve this problem for the advertising industry with SeekrAlign*, a solution that leverages proprietary scoring and analysis technology to help brands find the most suitable online content and grow their reach responsibly.

Built, trained, and deployed on an 8-node Intel® Gaudi® 2 AI accelerator cluster in the Intel® Developer Cloud, Seekr's first-to-market AI solution quickly garnered an impressive cross-industry customer list of leading brands. From Moderna and SimpliSafe to Babbel, companies are realizing the value of using AI solutions to solve enterprise challenges. Intel Developer Cloud's AI acceleration platform is at the forefront of this movement, enabling companies to accelerate and scale their AI strategies with cost-effective clusters of the latest Intel hardware and software.

Challenges & Opportunities of AI

Misinformation, hallucinations, and inappropriate content can quickly erode brand trust. Even inadvertently associating with platforms that host at-risk content can endanger a brand's reputation. For this reason, many companies closely screen the content they distribute. This practice is time-consuming, expensive, unreliable, and highly subjective.

A primary use case is podcasts. The popularity of podcasts has grown exponentially in recent years, and for many companies, podcasts are an essential part of their content strategy and advertising revenue. Unfortunately, the podcast landscape has been especially difficult for brands to navigate as, until now, there has been no reliable way to evaluate singular episodes other than human effort.

seekr™
seekr.com

SeekrAlign solves this challenge by allowing companies to quickly and accurately screen hundreds of thousands of podcasts. "Align simplifies finding safe and suitable podcasts by marrying world-class engineering with innovative and transparent ratings, allowing brands to navigate with nuance so they can confidently grow their audio campaign with increased clarity and speed," said Dan Granger, CEO of podcast advertising agency Oxford Road.[1]

Align's user-friendly dashboard rates podcast content on a scale between 0-100. Based on the GARM safety framework, this patented rating system, called the Seekr Civility Score™, gives marketers and advertisers an objective standard to judge content safety and suitability. "This challenge, at this vast scale, could only have been solved by combining superior computing power with novel signal detection," said Rob Clark, President and Chief Technology Officer of Seekr. "The Civility Score synthesizes dozens of considerations into a single metric that opens the door to more brand-suitable opportunities and larger audiences."[2]

The Align platform has already scored 20 million minutes of podcasts, a number that is projected to reach up to 40 million by the end of 2024. In addition, the platform has real-time scores for more than 8,000 podcast episodes, totaling more than 10 million minutes of audio. "We've shown that when AI is applied responsibly and transparently, it can solve real business challenges as this technology continues to develop and advance," said Clark. "I look forward to building on this success through continued innovation, including the development of advanced LLMs that can help solve some of the most vexing social issues and drive business growth with incredible speed and accuracy."[3]

Delivering High Performance with Reduced Cloud Compute Costs

Seekr's sophisticated AI workload requirements demand immense compute capacity. Historically, this compute power would require a significant infrastructure investment and considerable cloud costs. But with the Intel Developer Cloud, Seekr can access leading-edge Intel hardware and software for a fraction of the price. For a quickly scaling company such as Seekr, the ability to conserve capital and grow revenue while accessing the latest AI acceleration technology is an incredibly attractive proposition.

"This strategic collaboration with Intel allows Seekr to build foundation models at the best price and performance using a super-computer of 1,000s of the latest Intel Gaudi chip, all backed by high bandwidth interconnectivity," said Clark. Seekr's trustworthy AI products combined with the 'AI-first' Intel Developer Cloud reduces errors and bias, so organizations of all sizes can access reliable LLMs and foundation models to unlock productivity and fuel innovation, running on trusted hardware."

> "This strategic collaboration with Intel allows Seekr to build foundation models at the best price and performance using a super-computer of 1,000s of the latest Intel Gaudi chips..."

Flexible AI Solutions for Trustworthy Outcomes

SeekrFlow*, an end-to-end LLM development toolset focused on building principle aligned LLMs using scalable and composable pipelines, is the foundation of this innovative solution. Designed for maximum efficiency and flexibility, SeekrFlow allows developers to train and build LLMs using scalable and composable workflows. The entire LLM pipeline is managed with an API and can be adapted with custom components. Computational graph specifications such as resources and dependencies are flexible, allowing for countless configuration options.

SeekrFlow's workflow is compiled into a versioned, repeatable, and reproducible executable. It also integrates robust alignment algorithms to ensure the LLM's behavior complies with prescribed standards, intentions, rules, or values. These LLMs also benefit from the company's robust inference pipeline, enabling the platform to transcribe, diarize, and score audio at scale. Once optimized, these custom LLMs are piped into the user-friendly Seekr Align platform where they are used to provide fast and accurate evaluations of digital content.

Seekr Deployment in Intel Developer Cloud

Seekr had been using a regional colocation provider to host a server fleet of GPU and CPU-enabled systems. The self-managed compute infrastructure is used for LLM and multimodal model development and supports Seekr Align.

With customer growth and the increasing size of LLM deployment, Seekr sought a cloud services provider to help scale its businesses through superior price and performance. After successfully fine-tuning and inferencing benchmarking LLMs on Intel Gaudi 2, Seekr decided to leverage Intel Developer Cloud as its preferred computing infrastructure for LLM development and production deployment support.

Figures 1-3 show the performance gained through software optimizations and by using Intel Developer Cloud's environment.
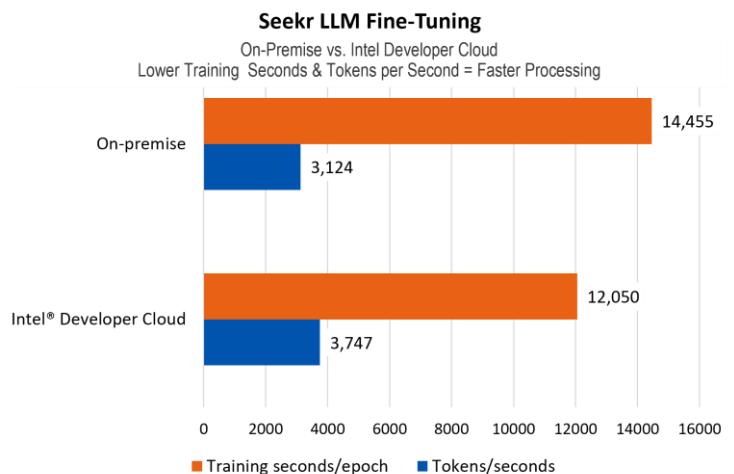
Seekr LLM Fine-Tuning
On-Premise vs. Intel Developer Cloud
Lower Training Seconds & Tokens per Second = Faster Processing

Figure 1: Seekr LLM fine-tuning using the Hugging Face Optimum Habana and Habana PyTorch APIs. Source: Seekr. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.
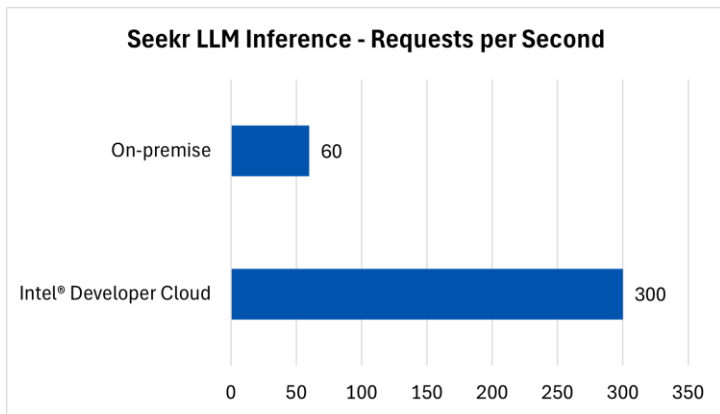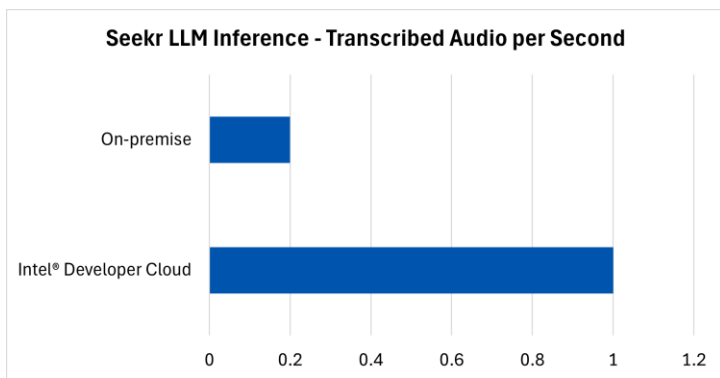
Seekr LLM Inference - Requests per Second

On-premise	60
Intel® Developer Cloud	300

0 50 100 150 200 250 300 350

Figure 2: Seekr LLM inference using Hugging Face Optimum Habana and Habana PyTorch APIs. Source: Seekr.
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Seekr LLM Inference - Transcribed Audio per Second

On-premise	0.2
Intel® Developer Cloud	1

0 0.2 0.4 0.6 0.8 1 1.2

Figure 3: Seekr LLM inference using Hugging Face Optimum Habana and Habana PyTorch APIs. Source: Seekr.
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Seekr plans a phased deployment to migrate the company's solution platform into Intel Developer Cloud. It is using various AI-optimized compute instances in our cloud to optimize the deployment cost/performance of application workloads, including:

- **Intel Gaudi 2 accelerators:** The 7B LLM model that powers the SeekrAlign solution is deployed on a cluster of Intel Gaudi 2 systems to perform sensitive content categorization and risk assessments.
- **Intel® Data Center GPU Max Series 1550:** Seekr uses Intel® Extensions for PyTorch* for transcription and diarization processing on a cluster of Intel GPU-powered systems.
- **Intel® Xeon® Scalable processors:** The company uses a cluster of 4th gen Intel® Xeon® CPU-enabled systems for application monitoring, storage management, and vector and inverted-index database services. These support Seekr's Retrieval Augmented Generation services and SeekrFlow on building **principle aligned** LLMs using scalable and composable pipelines.

Plans include utilizing 5th gen Intel® Xeon® Scalable processors later this year.

As Stefanos Poulis, Seekr's Chief of AI Research and Development, said "The Seekr-Intel collaboration unlocks the capabilities of running AI on hybrid compute. Intel GPUs and Intel Gaudi 2 chips are leveraged to train and serve at scale, large- and medium-size transformers. We saw improved performance for model training and inference compared to other chips in the market. Specifically on LLM inference, the large memory capacity of the Intel HPUs and GPUs has allowed us to adjust workload-specific parameters, such as batch sizes. On the software side, the various Intel extensions enabled us to move our ML stack to Intel hardware seamlessly."

As part of the following Intel Developer Cloud integration phase, Seekr plans to leverage large-scale Intel Gaudi 2 clusters with Intel® Kubernetes Service to train its LLM foundation model and add CPU and GPU compute capacity to grow its trustworthy AI service offering.

SimpliSafe® Reaches a Wider & More Tailored Home Security Audience Using Seekr's LLM Solution: SeekrAlign

SimpliSafe® is an award-winning home security solution on a mission to "make every home secure" by providing its customers with the latest and greatest technology. The company takes great pride in being a first mover in the ever-evolving smart home industry, and this sentiment is carried over to its supporting business functions, including testing emerging marketing technology that helps expand the company's market reach thoughtfully and efficiently. Additionally, building and maintaining customer trust is critical to SimpliSafe's customer expansion strategy. Not only does the company do this through its innovative products and services but also its thoughtful advertising practices.

SimpliSafe utilizes SeekrAlign to reach potential customers in the right moments through the right content channels. With SeekrAlign, SimpliSafe has access to content scores for thousands of podcasts in real-time and can isolate context at the conversation level. Seekr's LLM-driven content analysis abilities go beyond anything available on the market today and help SimpliSafe discover brand-suitable content to reach relevant audiences.

"As a brand driven by trust and safety, it was important for us to find a solution that shared SimpliSafe's values. Seekr powered by Intel Developer Cloud is just that platform as it empowers us to wrangle the scale problem of monitoring podcasts thoughtfully and ensure that where we advertise matches our customer profiles. With Seekr, we are able to find brand-safe podcasts through which we can reach the right audiences and ultimately introduce SimpliSafe's home security solutions to more homeowners."

Nicholas Giorgio, Director of Customer Acquisition, SimpliSafe

More Information
    Seekr – www.seekr.com
    Intel® Developer Cloud – cloud.intel.com
    Intel® Liftoff – developer.intel.com/liftoff
    Intel Gaudi 2 AI accelerator – habana.ai
    Intel Enterprise Software Solutions

About Intel® Liftoff for Startups Program
Intel Liftoff is a free virtual program that helps early-stage AI and machine learning startups innovate and scale. Launched in 2022 by Intel, this self-paced program empowers startups to remove code barriers, unleash performance, and transform their ideas into scalable, industry-defining AI companies.