INFERENCE: WHERE AI TRAINING ENDS & BUSINESS BEGINS

Key Enterprise Considerations
for Inference Deployment of Large Language Models

groq™

Executive Summary

Historically, most investments in artificial intelligence (AI) systems have focused on training. We are now at an inflection point where business leaders need to move their trained AI models into production, to inference. Inference uses input data to solve real-world challenges, enabling businesses to compete in a market abundant with data that demands real-time insights at an accelerated time-to-production. Training–deriving insight from data–was the necessary first investment when building an AI strategy while inference turns that data into profit by operationalizing production-ready workloads and models to help with real-world and real-time decision making.

There are a wide variety of applications that sit under inference, including computer vision, computational sciences, linear algebra, real-time series, anomaly detection, and most notably from a public attention standpoint, Natural Language Processing (NLP). **One of the leading inference applications under NLP is Large Language Models (LLMs)**. In this paper, we will share the four fundamental considerations AI business leaders should look at when evaluating their inference strategy for LLMs.

KEY TAKEAWAYS

1. **Pace:** Move at the Rate of Innovation

The best AI inference strategy accelerates LLM model deployment and enables the flexibility for developers to quickly adapt new model architectures to their custom software solutions.

2. **Predictability:** Know Performance With Clarity

Determinism, a system's ability to deliver predictable and repeatable performance, is the only way to 100% consistently guarantee critical performance metrics such as throughput, latency, accuracy, and power consumption.

3. **Performance:** Achieve Faster Token Output

When evaluating inference processors for deploying autoregressive LLMs, it is crucial to consider the rate of tokens output per second, not just the rate of tokens input and processed per second.

4. **Accuracy:** Get Smarter With Every Prediction

Improving LLM accuracy and avoiding hallucination through techniques such as reflection is critical to both business outcomes and overall resource consumption.

We will also share important questions for leaders to ask potential inference partners regarding LLM inference solutions. Throughout the paper, you will learn more about the LLM inference landscape, including high level market data points, the growing number of emerging enterprise applications, and possible deployment obstacles to consider. We will also introduce Groq, our technology, and the advantages our software and hardware solution ecosystem offer for inference and specifically LLMs.

groq™

Introduction

Business leaders developing an AI strategy have a challenge. Training, the development phase where a new model learns to make predictions by studying large data sets, is very expensive. Making good on this investment requires moving from training to inference, the deployment phase where a trained model generates real-time predictions based on new data. Inference is where AI workloads start to earn their keep.

The challenge is, GPUs, which dominate the AI hardware world today, are okay at inference for some workloads but fully unequipped for others. These "others" include workloads that demand real-time performance, such as LLMs that many organizations are racing to get into market. In GPU-dependent AI systems, inference performance on these types of workloads often isn't optimal enough to justify the high costs, threatening the viability of many business cases.

The Rapid Rise of NLP

According to Mordor Intelligence, the 2023 NLP market size is estimated at **$25.62B USD**, is expected to reach **$75.01B** by 2028, and is growing at a **CAGR of 23.97%** during the forecasted period of 2023 to 2028.[1]

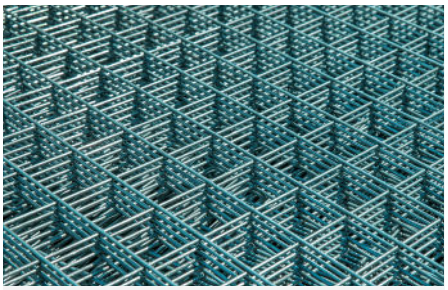
To be successful, AI business leaders need real-time, highly accurate insights at the performance and price point that supports their business case at scale.

For AI business leaders, following a sound inference strategy will be the difference between success and failure when it comes to deploying LLM workloads. To be successful, AI business leaders need real-time, highly accurate insights at the performance and price point that supports their business case at scale. Get the inference strategy right, and enterprises can achieve a generational leap in the return on investment (ROI) of AI solutions, whether they're leveraging LLMs or some other revolutionary workload. Get it wrong, and AI will continue to sit in research and development as a cash burning project, not economically viable or imperative to the business.

At Groq, we believe there are four factors AI business leaders should consider when evaluating their inference strategy: **Pace**, **Predictability**, **Performance**, **and Accuracy.** This paper is a deeper dive into those considerations, as well as important questions for leaders to ask potential inference partners.

Pace: Move at the Rate of Innovation

The pace of LLM innovation is hardware-constrained. New model architectures are constantly emerging, both to push the limits of what AI can achieve and to simplify and lower the costs of scaling. Meanwhile, there are new workloads everyday given the low barrier to entry of software compared to hardware, along with AI bots playing a role in generating and developing new algorithms. These fast-moving software trends are outpacing hardware capabilities. When training a model is completed, it takes many months and seemingly endless resources to navigate a fragmented ecosystem, create custom kernels (software routines for data transformation and mapping), and compile and deploy the new workload. Managing this process requires a specialized skill set that doesn't exist in many enterprises today, and one that can be challenging to hire.



The Growing Use of Generative Models

While 41% of decision-makers planned to deploy generative models this year, 28% stated their infrastructure doesn't currently meet their technical requirement. Appropriate investment in inference solutions are needed to address this lag.[2]

The best AI inference strategy accelerates LLM model deployment and enables the flexibility for developers to quickly adapt new model architectures to their custom software solutions.

The best AI inference strategy accelerates LLM model deployment and enables the flexibility for developers to quickly adapt new model architectures to their custom software solutions.

Questions AI business leaders should consider when planning their inference strategy:

- Will we be able to move at the pace of software innovation?
- What is the timing, complexity, and cost to move a workload from training to inference?
- What are my human capital requirements? What type of people? How many of them?

Predictability: Know Performance With Clarity

When designing new AI solutions, developers have limited line of sight into how their workload will run once it is actually deployed and scaled on hardware. **Critical performance metrics such as throughput, latency, accuracy, and power consumption can be estimated but only with determinism–a system's ability to deliver predictable and repeatable performance–can they be guaranteed with 100% accuracy.** Developers are forced to deploy their workloads, measure, adjust, and re-deploy, time and time again, wasting time, money, energy, and silicon.

As a result, LLM developers lack clarity on a workload's performance and cost. This can make it challenging to optimize software design for a particular performance metric given the time gap between making a tweak and re-deploying.

Questions AI business leaders should consider when planning their inference strategy:

- Will we understand how a workload will perform with exact clarity? If not, what trade-offs are we making in regard to time, money, and resources?
- Will we be able to co-optimize software and hardware to get the trade-offs just right?

Potential Pain Points for Enterprise Deployment of LLMs

While determinism will help with predictability of critical performance metrics, there are other deployment pain points to be aware of. Is your AI strategy equipped to navigate the following?

DEPLOYMENT AT SCALE

DATA PRIVACY & IP PROTECTION
Especially when using 3rd party models and infrastructure

HUMAN CAPITAL
Required to operationalize LLMs

INFRASTRUCTURE COST & AVAILABILITY
For both compute hardware and supporting resources (e.g. power)

Performance: Achieve Faster Token Output

If you want to grasp the importance of LLM latency–a key component of performance–try this with ChatGPT. Cut and paste a thousand word block of text into the query box, preceded by the request, "Summarize this passage in 100 words." The model will return the first word of its response in under a second, but it will take about nine seconds to complete its 100 word answer. GPUs are relatively good at input, the initial reading and processing of data that occurs in a split second, which is great when training a model.

But when running inference for an LLM model, output speed becomes much more important. What matters with generative AI is the generative part: the ability to create a response by inferring a word–a token in AI parlance–then the next one and the next one. GPUs falter when it comes to running this type of sequential (autoregressive) AI model because it requires frequent and rapid memory cache updates. This entails a lot of back and forth to off-chip DRAM, hence the nine seconds it takes for ChatGPT to generate a 100 word answer. GPUs deploying LLMs are like a human–they can read and process information much faster than they can generate it.

When evaluating inference processors for deploying autoregressive LLMs, it is important to consider the rate of tokens output per second, not just the rate of tokens input and processed per second.

Questions AI business leaders should consider when planning their inference strategy:

- How fast will my model run?
- What will its latency be?
- At what sequence length?
- At what cost and power consumption?

Emerging Enterprise LLM Applications Across Verticals

Each application across verticals has unique performance requirements. Which ones are critical to your organization now, and which ones do you plan to invest in over the short and long term? Are your AI systems capable of delivering a performance standard that universally serves all of your applications?

FINANCIAL SERVICES
- Risk Assessment & Fraud Detection
- Market Research & Competitive Analysis
- Regulatory Compliance
- Customer Support & Service

GOVERNMENT & PUBLIC SECTOR
- Policy Content Generation
- Text Summarization & Document Analysis
- Data Analysis & Insights
- Chatbots & Virtual Assistants

RESEARCH & SCIENCES
- Data Analysis & Insights
- Text Summarization & Document Analysis
- Knowledge Management

CYBERSECURITY & INFOSEC
- Risk Assessment & Anamoly Detection
- Legal & Compliance
- Sentiment Analysis & Monitoring

ENTERPRISE COMMUNICATIONS
- Voice Assistants & Voice Interfaces
- Language Translation & Localization
- Customer Support & Service

Accuracy:
Get Smarter With Every Prediction

The end objective of any LLM workload is to deliver the most accurate output possible, whether it be the next word, line of code, or market insight. Training enables this, giving AI workloads a sort of intuition to look at a situation and be able predict what should happen next (e.g. the next word in a sequence) based on how it was trained.

When training is complete, learning is not. There are several techniques developers can employ during inference to improve accuracy so the model can continuously learn and get smarter. A sound inference strategy must be able to support this capability.

Some LLMs use a type of search-ahead function during inference to "test" their intuitive-like answers (again, based on their training) by searching ahead multiple steps to see if a better option emerges. For example, the Go-playing AI programs AlphaGo and AlphaGo Zero, developed by Alphabet's DeepMind group, used this technique to achieve much higher player ratings, or their Elo score (read more here). Another approach is to use an algorithm called beam search, which is similar to the search-ahead technique employed by AlphaGo. These techniques are fading in popularity, though. As training gets better, the models have gotten smart enough that their initial prediction is quite good, rendering techniques such as beam search less helpful. Seeing a few tokens (or moves) ahead isn't worth the expense.

Where Does Your AI Maturity Stand?

C-suite members are nearly 2x more optimistic than highly technical roles to perceive their AI maturity as "very mature." Using the questions provided in this paper can help facilitate conversations to form a unified perspective on AI maturity at an organizational level and identify opportunities for advancement.[3]

As training gets better and search-ahead techniques less effective, other techniques are emerging to enable continuous learning. For example, an approach called "reflection" waits until the output is complete and then looks back at the outcome and evaluates how it could get better. This is akin to how a human might approach writing an article or story. They use their training (both classroom learning and experience) to write their first draft, then go back and edit (and edit, and edit) what they have written. With each pass they ask themselves, how can I make this better?

ChatGPT's responses are essentially first drafts; it has no ability to go back and edit itself as a human would. But what if it did? Try it. The next time you submit a query and get a response, follow up by asking it to "list five ways your answer could have been better, pick the top options, and apply them to the answer." This usually directs the bot to provide a higher quality response.

These techniques to improve accuracy during inference take time, so using them effectively requires low latency. When considering LLM inference strategies, AI business leaders should consider how they will start to use techniques like reflection to improve their LLM's accuracy, and how that will affect both their business outcomes and overall resource consumption.

Questions AI business leaders should consider when planning their inference strategy:

- How can my workloads get smarter during inference?
- Do I have low enough inference latency to support these solutions?

Groq, Your LLM Inference Partner

Up until recently, the compute industry was hardware-led with software following suit. Over time though, accessibility to software development became ubiquitous, snowballing into the explosion of software we see today. At Groq, we anticipated that this exponential growth of AI models would require an equally exponential amount of human resources in the form of developer hours. With this unsustainable future on the horizon, we got to work on this problem from a first principles approach.

We first built an easy-to-use software suite and then designed a low latency hardware architecture, called the Tensor Streaming Processor (TSP), purpose-built for AI. Let's get into the specifics of why Groq is your inference partner when you are ready to take off the LLM training wheels.

Pace: Move at the Rate of Innovation

Groq solutions provide faster time-to-market for LLM workloads with far less complexity and cost. Our kernel-less compiler can process most workloads in a small fraction of the time of GPU-based inference systems—days not months—and requires far fewer engineers. This not only accelerates the pace of solution development and deployment, it solves the human capital problem. **With Groq, you need fewer people to deploy and scale LLM workloads.** For example, Groq's development team of ~40 engineers compiled over 500 models to run on its hardware in just 45 days.

Predictability: Know Performance With Clarity

With Groq, developers can clearly understand what software can expect from hardware early on in the development process. Specifically, **Groq solutions, including our software tools and deterministic architecture, enable developers to optimize instead of making the historical trade-offs between hardware and software, and produce predictable and repeatable performance metrics at scale.** With this, they can confidently look ahead, and identify and adjust performance-impacting components of their model.

Performance: Achieve Faster Token Output

For large scale LLMs, Groq is simply faster at inference. How much faster depends on many factors, but in many cases it's more than 10X. **The question then becomes, where would you like to activate that 10X improvement?** Throughput? Cost savings? A balance of both? Groq helps its partners answer this question to achieve their goals.

Accuracy: Get Smarter With Every Prediction

Our performance advantage gives developers the opportunity to improve inference for an LLM workload's predictions in real-time, using reflection or similar techniques. More accurate insights, better business outcomes, a stronger ROI–all possible when using Groq for inference.

We'd love to partner with you for your LLM application deployment, and know we can do it because of some of the key advantages we offer.

Better Performance at Scale

Benefit: Comparable LLM throughput vs Nvidia A100 DGX solution with 2-5X upside*

Push Button Compilation

Benefit: No model sharding or waiting for vendor to release compilers / kernels for optimized model

Lower Power Consumption

Benefit: Up to 2-3X lower node max power consumption (GroqNode™ 1 vs Nvidia H/A100 DGX)

Lower Latency

Benefit: Extremely low latency modes with 5–10X advantage over Nvidia A100 DGX*

Product Availability

Benefit: Groq products available for purchase and deployment in scaled systems

Domestic Supply Chain

Benefit: Groq hardware is designed, manufactured, and assembled in North America while Groq software is sourced in North America

*Based on Groq projected performance, subject to change

Conclusion

As the inference leader, Groq talks with a lot of enterprises who are trying to determine the optimal inference strategy for their enterprises. The most common question we hear is, “What will our TCO be with your solutions?”

It’s a good and valid question, but not sufficient. Here are some additional queries to include:

- **Will we be able to move at the pace of innovation? What do we need to get there?**
- **What is the timing, complexity, and cost to move a workload from training to inference?**
- **What type of people do I need? How many of them?**
- **Will we understand how the workload will perform with exact clarity?**
- **Will we be able to co-optimize software and hardware to get the trade-offs just right?**
- **Will my model run fast enough? What will its latency be? At what cost and power consumption?**
- **How can my workloads get smarter during inference? At what cost and power consumption?**

LEARN MORE

When considering these questions, Groq is prepared to help move your organization from training investment to inference value.

If you’re interested in learning more about our solutions or connecting with a Groq specialist, please reach out to us at **contact@groq.com**.

ADDITIONAL READING

AlphaGo & Learning During Inference

To understand how developers can improve the accuracy of their models during inference, it helps to look at the history of Deep Mind's AlphaGo, a generative AI system for playing the board game Go. The first iteration of AlphaGo was trained using data from thousands of games. This created a model with a seemingly intuitive ability to play Go: when presented with a new game situation, it knew what its next move should be.

But when playing the game, it didn't just rely on its training. Once it had determined what its next move would be, it used a type of search algorithm to look ahead and play out what the opponent's move would be, then its response, and so on. In fact, it did this for not just the next best predicted move, but for a list of the next best N predicted moves, then N counter-moves and so on, where N was based on factors such as compute constraints. For example, the model would predict the top 10 best next moves to make, then what the 10 best counter-moves would be, and so on. It analyzed all these branches to figure out which move it should make. It used this technique to test how good its initial move was.

This is the AlphaGo version that beat European champion Fan Hui in October 2015. It required 176 GPUs and achieved an Elo score (a mathematical rating of a player's performance) of 3,144. After that match, Deep Mind moved AlphaGo inference from GPUs to the more powerful TPUs. This enabled a more robust search-ahead capability, so the same model, with the same level of training, could make smarter moves. The result: running on only 48 TPUs, AlphaGo beat world champion Lee Sedol a few months later and achieved an Elo score of 3,739.

It quickly got better, so much so that it came to rely on its search-ahead capability less and less.

This approach generated some creative play. In particular, in its second game against Lee Sedol AlphaGo made a move (called a "shoulder-hit") that wowed many top players as being especially unusual and creative, something that might appear in only 1 in 10,000 games.

The next iteration of AlphaGo, called AlphaGo Zero, was trained in an entirely different way. Rather than learning from a vast library of historical games, it learned by playing itself (reinforcement learning). At first it was terrible, since it knew absolutely nothing about Go. Searching ahead helped it play better, and everytime it concluded a game against itself it added that game to its training database. It quickly got better, so much so that it came to rely on its search-ahead capability less and less. Searching is the most compute-intensive component of running AlphaGo, as players have about 250 possible moves every turn, so relying less on search and more on initial prediction meant that AlphaGo Zero was much more efficient. In fact, this approach resulted in an Elo rating of 5,185 and a 30% improvement in outcome at a 92% reduction in hardware cost.

Citations

1 The Rapid Rise of NLP

According to Mordor Intelligence, the 2023 NLP market size is estimated at $25.62B USD, is expected to reach $75.01B by 2028, and is growing at a CAGR of 23.97% during the forecasted period of 2023 to 2028.

https://www.mordorintelligence.com/industry-reports/natural-language-processing-market

2 The Growing Use of Generative Models

C-suite members are nearly 2x more optimistic than highly technical roles to perceive their AI maturity as "very mature." Using the questions provided in this paper can help facilitate conversations to form a unified perspective on AI maturity at an organizational level and identify opportunities for advancement.

Source: According to a 2022 Deloitte survey focused on 600+ global, cross-industry decision-makers from the fields of AI, analytics and data. https://www2.deloitte.com/uk/en/insights/focus/cognitive-technologies/ai-and-machine-learning.html

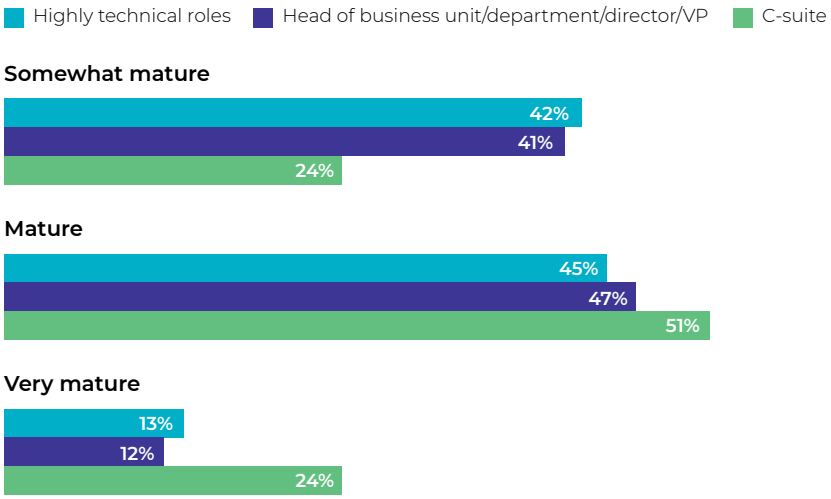
3 Where Does Your AI Maturity Stand?

When considering AI-maturity perception, C-suite members are more optimistic while highly technical roles are the least likely to view their organization as mature or very mature. Using the questions provided in this paper can help facilitate conversations to form a unified perspective on AI maturity at an organizational level and identify opportunities for advancement.

Source: According to a 2022 Deloitte survey focused on 600+ global, cross-industry decision-makers from the fields of AI, analytics and data. https://www2.deloitte.com/uk/en/insights/focus/cognitive-technologies/ai-and-machine-learning.html

C-suite are the most ambitious about current AI maturity

Question asked: How mature would you rate your organisation in terms of AI?



Source: Deloitte analysis

INFERENCE: WHERE AI TRAINING ENDS & BUSINESS BEGINS

AUTHOR

Alan Eagle

Author, executive communications coach, former Google Managing Director, alaneagle.com

Acknowledgements: Niamh Gavin, Mark Heaps, Mike Henry, Brian Kurtz, Mariah Larwood, Andrew Ling, Jake Louderback, Jonathan Ross, Ramakrishnan Sivakumar, Graham Steele

groq™