Machine Learning Engineer Assessment Test

***Candidate's Name*: Ravali Yerrapothu**

PART ONE

1. Are you currently employed?

No

2. How many hours can you devote weekly to the Company?

40

3. Why do you think you fit the role well? Please provide examples of your skills and experiences that align with the requirements of this role.

I am an experienced ML engineer with experience in vast areas like traditional ML, neural networks and generative AI and with a background in Computer Science.

My educational background which is Masters in CS with focus in Machine Learning and Expert Certificate in Deep Learning from CMU are a testament to my continued interest and expertise in the field. All of which are directly relevant to the skills needed in this position.

Some of my project highlights are Training LLMfrom scratch, Custom libraries for Neural networks, Generative AI base summarization and agentic ai based scientific IDE.

4. Have you worked from home before?

YES

5. What interests you about this role?

Working on cutting edge ML problems that impact humanity.

Getting an opportunity to work on deep neural networks and research and translation of the features to products.

Working alongside some of the pioneers in the industry and being able to learn from them is invaluable.

6. What motivates you?

Advancements in Machine Learning, technology and its impact on people like reading about nice problems engineering are solving for people with novel products, curiosity, and zeal to contribute /be helpful to people.

7. Are you seeking employment in a company of a specific size, such as a small startup, medium-sized Company, or large corporation? Please explain your preference.

Preferred - Mid size, Small startup. I want to be part of a small and supportive team that makes valuable contributions to the product.

I have prior experience in startups and small sized companies <50 employees, and large sized organizations. I felt that small companies will be able to provide a platform where we get to work on exciting features at a rapid pace and teams stay motivated because of the impact and belonging.

8. Describe your experience working with a culturally diverse group of people

- While working on Google Ads for Google via Virtusa, the team was diversely located in geography and so is culture. Many of the engineers were working from Mountain view CA, and New York, and few in Europe. But most of them that I interacted regularly were American Nationals, and South Asian nationals primarily India and China.
- Our own team from Virtusa had people from several areas in Northern and Southern India, each of which has a peculiar background.
- During my time at SIOS Technology, I got to work with engineers from South Carolina with different backgrounds in CPP, Java and Statistics.
- During my Master's I interacted and worked as a team with people from Nepal,and Pakistan.

9. - What are the primary differences between transformer architectures and RNNs? When would you choose each, considering computational efficiency and performance?

RNNs use recurrence at a layer as a hidden state from the previous layer where all information is compressed. Forms of RNN like LSTM, GRU, use different forms of recurrence and gates to extract specific information. There is no means to check how the words in a sentence are related.

Recurrence is a hindrance for parallelization during training.

These may be used when data is small or memory is constrained.

Also in situations where we need low latency during inference time.

In Transformers, The attention module in that Self Attention -> specifies how the words in the sentence are related, This provides the context. In addition to this, Positional embedding specifies the exact position of the token.

There are several forms of attention like multi head attention, masked multi head attention. which have different techniques underneath.

Can be fully parallelized.

Computation is expensive compared to RNNs but may take advantage of parallelization.

- How would you design a model for long-range dependencies that also meets real-time inference constraints?

Despite variants like LSTM/GRU, long range dependencies are a problem for RNNs.

Transformers have direct access to all tokens. But they may get expensive. Hence variants like FlashAttention may be used.

10. - You're training a large language model and notice that the loss is decreasing, but validation performance has plateaued. Walk me through your debugging process and the potential causes you would investigate.

Training loss is decreasing and validation loss is plateauing, which implies a classic case of the model overfitting to the data:

I would start by inspecting the data, hyperparameters, model architecture and evaluation strategies step by step.Check training and validation data distribution and adjust for any differences.If more data is available, I will add more higher quality data as the model may have extracted all useful signals. Check for any leakages in training.I try to employ cross validation and add further splits and measure the training loss over different folds of data which gives a more accurate estimate.

Regularization and Hyperparameters: I will add regularization parameters to reduce over estimation of weights. Add Dropout or attention dropout, Use Early Stopping and Reduce number of training steps. Check LR schedule and decay. Employ optimization like AdamW, and Gradient Clipping thresholds.

Verify model architecture & revisit model size, layers, nodes, attention heads etc., Evaluation of metrics must be thorough and specific to the task.

- To clarify, could you describe your approach to efficiently implementing distributed training across multiple GPUs or nodes? Also, what specific challenges have you encountered with gradient synchronization?

My goal while implementing distributed training is to maximise hardware utilization and minimize communication overhead.

Choosing right parallelism strategy Data Parallelism vs model parallelism.In Data parallelism each GPU has a full model replica and batches are split, Uses Distributed Data Parallel (DDP). Gradients are synchronized via AllReduce. This is employed when the model fits in one GPU. Eg: fine-tuning or medium sized LLMs. Used with large batch sizes. Model Parallelism: Done in 2 ways Tensor parallelism where we split large matrix multiplications across GPUs. Pipeline parallelism where layers are split across GPUs. Uses micro batching to keep pipelines full. This is typically used in LLM pretraining.

Furthermore, Sharded approaches like FSDP or Deepseek ZERO where parameters and gradients are sharded across GPUs. Mixed precision training: Using FP16 and BF16 to reduce weight storage. Scaling checks: I track metrics like throughput: tokens/sec, gpu utilization, communication vs compute ratio

The most common gradient sync issues I've seen are communication bottlenecks, deadlocks from rank mismatches, numerical instability with mixed precision, and scaling inefficiencies due to batch-size changes. I usually debug by profiling backward vs AllReduce time, validating identical execution across ranks, and sanity-checking gradients before and after synchronization.

PART TWO

1- Describe the differences between Adam, AdamW, and SGD with momentum. When would you choose each optimizer, and how would you tune its hyperparameters?

SGD, Adam and AdamW are all gradient optimizer techniques to reach convergence. SGD updates gradients by randomly selecting a sample from the training set, estimating error at that point and shrinking LR. This can be noisy. It suffers from higher variation between runs due to randomization, converges quickly but to worse loss than batch descent. Momentum helps smooth out the noise by using the running average of the gradients rather than raw values. This smooths out the oscillations. But it may still result in exploding or vanishing GD.

Adam is a higher order optimizer, combining RMSProp with and momentum, where gradient is estimated for all parameters and is estimated as root mean square of gradient for that parameter. Adam maintains running averages of mean derivative mk and mean squared derivative vk. Applies L2 regularization.

AdamW is an improvement to Adam which adds weight decay by decoupling it from adaptive gradient updates. Instead of L2 regularization, it applied weight decay directly to weights, without interacting with LR updates, making it more effective with improved generalization.

- What regularization techniques would you use to prevent overfitting in an extensive neural network, and how do their mechanisms differ?

- Use L1 and L2 regularization techniques to reduce over estimation of weights and employ Weight Decay.
- Adding Dropout and other forms of Dropout like Zoneout for RNN, DropConnect where we drop individual connections instead of nodes., Shakeout: Scaling up randomly selected weights and fixes rest and Whiteout: Add or multiplies a weight dependent Gaussian noise to the signal on each connection.
- Use Early Stopping, Gradient Clipping and reduce the number of training steps
- Normalizing the data, and using initialization methods like Xavier and Kaiming.
- Data augmentation methods when there is not enough data, by adding or changing data like rotating, translating, stretching, adding noise to images such that it doesn't change their class.
- Check LR schedule and decay: Employing techniques like linear decay, quadratic decay and exponential decay.
- Employ optimization like AdamW, and Gradient Clipping thresholds.

2: Share a challenging technical issue you've encountered and how you resolved it, focusing on your analytical and critical thinking skills.

During my initial days when I was working on problems in Computer Vision and even now when I am trying to understand an obscure concept like Transformers or for example GRU, I have found that some of these very tough concepts have answers in simple intuitions and that helps in solving many problems in terms of how they work. The same applies to any machine learning challenges.

Multi Agent framework for Seasonality.

The Seasonality implementation was adapted as a multi-agent framework, where each agent named as Daily Agent works on extracting daily trends, Weekly Agent provides Weekly patterns, and monthly and yearly agents work and extract monthly and yearly trends appropriately. Decisions like sample size, sampling rate, distribution, parameter tuning during modelling for each seasonal agent, Integrating the results from all the agents, the design and implementation of the components in software - classes, threads, interactions, asynchronous executions, and write unit and integration tests, was challenging engineering feature I have worked on.

3. How do you handle feedback and criticism from supervisors or colleagues?

I handle criticism and feedback with both positive and negative feedback well. I tend to introspect myself and constantly try to look for ways to correct and improve myself and my work. For this reason, I am usually aware and inline with the feedback by the time it is given to me.

4. What is your greatest strength? How will it help you in this role?

My ability to comprehend complicated concepts and effectively solve and implement them is one of my greatest strengths. Along this way, my persistence, and attention to detail, helps in staying focused and consistent.

I think this role requires us as Engineers to work on some of the very challenging problems in terms of Infra development for data, model or inference pipelines, training the models and optimizing them with GPUs, all of which require us to be very detail oriented and requires us to be persistent in hanging around during long training cycles fixing any errors we made along and reiterating for optimizations etc. and for us to understand and improve the very exciting model architectures etc.,

5. We want to understand your long-term career goals and how this position aligns with them. Please share your aspirations and how you see this role contributing to your career path.

I want to grow technically and progress to a lead and architect ML product development. In that, I want to gain the skills from best of both the worlds ML research and feature development and this position is exactly that where I get to learn from and contribute to frontier ML research and products including design and implementation of large scale machine learning systems and infrastructure including training and inference pipelines, translating cutting edge research ideas to production, model evaluation and optimization with safety and alignment in focus.

I truly believe that in spite of the recent growing trend of rapid productionization of research ideas, there is a lot of research which needs to be brought to people, in forms of useful products., and I want to do that in the best possible way.

6. What do you know about our Company?

I got to know about Anthropic in 2023, when I was working on an LLM powered project for summarization, and was pretty amazed at how well the responses from Claude were. Since then I have been following the models (Haiku, Sonet, Opus), products (Claude, Claude code) and community developments including MCP, Agent Skills from Anthropic.

Anthropic's main mission is to promote AI research and development that promotes AI safety and alignment.

Anthropic has been one of my dream companies to work at and it will be my honor if I get the chance.

7. How do you see yourself contributing to our Company's continued success and growth in this role?

- Adding my contributions to some of the challenging problems into the product.
- Crafting artifacts and technical papers, publishing white papers and engaging in conferences representing the organization.
- Mentoring juniors and contributing to knowledge sharing.
- Bringing my diverse cultural and educational background and ideas that stem from it.

8. In your opinion, what are the biggest challenges facing our Company in the market?

The AI market has become very competitive with any and many products launching every day.

The talent shortage is real, as AI talent is highly sought after in the current market.

The costs for maintaining the product infrastructure, training and inference pipelines is bulky in itself and adding to it is the compensation for the engineers.

Also navigating regulatory, data privacy, licensing issues, gaining trust are some of the ethical and moral challenges.

9. Discuss your familiarity with recent technological trends and how they can be leveraged within the industry.

I ensure that I regularly upgrade my skil set in accordance with industry trend, This is also inline with my personal trait of being curious. And so it helps me stay on toes.
I specialize in Deep Learning from CMU.
My masters has been focus in Machine Learning, Data Mining and Computer Vision.
I have industry experience in traditional ml, and generative ai.
I participate regularly in community events like ODSC conferences and meetups.
Recently, securing second place in a hackathon conducted by Google Cloud.
This all round skills helps me adapt quickly and pick any task needed and run with it from initiation to completion. My varied skills helped understand in depth, and what happens behind the hood in systems, ml models, and mathematical understanding. This skills I believe are very unique and must be taken good care of.

10 - What do you see as the most important open problems in AI safety and alignment? How would you prioritize research efforts in this space?

The key open problems in AI safety are value alignment, robustness under uncertainty, interpretability, scalable oversight, and predicting long-term societal impacts. Research should prioritize immediate robustness and interpretability, followed by value alignment, and finally long-term forecasting to ensure AI systems act safely, reliably, and in line with human intentions.

PART THREE

1. What tools or strategies are most effective for communication and collaboration in a remote work environment?

    Staying connected via chat or quick meetings will help us resolve any hindrances the team is facing that is slowing down the development.
    Effective email updates will keep everyone posted.
    Regular team meetings will improve connection between teams and improve collaboration. It also helps in assigning work items for the day, in addition to sprint meetings.
    Tools: Google meet is very quick and robust,other video calling tools like zoom may also be used, slack for team chat, google calendar.

2. If you were hiring for this role, which key personality traits and qualities would you look for?

    I will look for a person who is an Engineer at heart - who likes to explore, learn and build things.

    Most of the problems encountered here do not have clear text book solutions, and require creativity and exploration and curating solutions. Learning is a never ending process, and quick learning is a must.

3. How do you approach problem-solving as a team member versus an individual contributor?

As a team member I like to throw my ideas and solutions out there to discuss with other people and get their inputs. We would discuss pros and cons of each approach and finalize the best solution.

As an individual contributor, I note down the problem,break it into smaller sub problems, research and gather solutions, and list pros and cons. I try to deduce the best approach using my knowledge, or researching further if I am short of.

4. How do you prioritize and manage your workload in a fast-paced environment?

I attend this by employing prioritization, planning, proactive communication and disciplined execution.

- I tend to make plan by scheduling things, in terms of priority, complexity, time needed to complete and whether it requires deep focus etc., Making a todo list helps me stay on track in terms of completing a higher number of tasks than otherwise and helps in grouping similar things together, so they can be completed even quickly.
- If time is short, I tend to do this mentally, and I review and adjust the plan continuously based on progress.
- If any miscellaneous tasks of lower priority can be skipped and revisited later, I make a note, and start jumping into critical things.
- It is of utmost importance to proactively communicate regarding any issues, blocks or hindrances to resolve them quickly and proceed to completion.
- I stay disciplined and focused in executing the plan, reiterating and working on things in order.

5. What specific skills or experiences do you hope to gain from this role to help you achieve your career goals?

Specific skills I am looking to gain are:

- Design and development of large scale world class machine learning systems.
- Improve my understanding of novel architectures in research for transformers, and large models including State Space, diffusion based models, etc.,
- Building tools and frameworks that accelerate research and collaborate and translate cutting edge ideas into production ready systems.
- Distributed training and large scale ml infra development,
- Optimization of ML training and inference pipelines including GPU based parallelization, quantization, inference acceleration etc.,
- Designing benchmarks and frameworks for model evaluation, alignment, monitoring and interpretability.

6. Imagine that you told a client you would be there at 10 am. It is now 10:30 am, and you won't finish your job until 11:30. You have a lunch meeting with another client at noon, followed by another job at 1:15 pm. How would you handle this situation?

I would handle this by being proactive, transparent and respectful of everyone’s time.

First, as soon as I realize that it is 10:30 and we are past 10 AM, I will contact the 10 am client and explain the delay briefly, take responsibility and explain the new job completion time i.e, 11: 30 am.

If the client agrees, and the meeting agenda can be accomplished within the new time slot, I try to finish my job and go for the meeting. I would also inform the client scheduled for the lunch meeting that I may be a little late, so that they can plan accordingly.

If my job is of less priority and can be paused and resumed later during the day or next day morning, I would prompt, pause and rush for the meeting, again doing my due diligence and informing both the clients about the delay and schedule changes.

7. In your experience, how frequently does the following problem occur: employees being afraid to express disagreement with their managers?

I faced such a situation only once in one of the projects at Virtusa. Even then, the manager was open to suggestions to improve the operational ways as there was clear friction in the project cycle. But I heard about it a lot from peers. But I believe it would come down to 1 or 2 on 10. As there is a clear problem in hiring if a person is not in alignment with the manager.

8. Your experience working on projects involving a consortium of companies is valuable. Please briefly share your experience in this area.

I worked briefly on SWE Bench and Data Science Expert trials for Mercor, which is an AI based organization that hires people primarily for AI benchmark evaluation projects for several large companies.

9. How much would you request per hour?

60$

10. When would you be available to start if hired?

Immediately.