Matthew Virginia
CS 5393
10th April, 2025

Exam 2

Model Overview

I chose three open-source language models for this project: TinyLlama, Llama 3.2, and Deepseek-R1:7B. I aimed to compare an array of options that varied in size, resource requirements, and popularity among open-source communities. My intention was to probe how each of them would function in real-world HR-oriented tasks like authoring job descriptions, coming up with interview questions, and assessing candidate answers. I made the choice deliberately with every model contributing something distinct, hoping it would result in more informed comparisons and richer insights.

I opted for TinyLlama mostly due to its size and efficiency. It's a light model, and I was interested in how far a scaled down LLM would be able to go in terms of providing useful, high quality outputs. Because HR workflows may involve repetitive or resource-intensive tasks, a light machine like TinyLlama could potentially be an affordable solution. Adding it allowed me to experiment with the practical trade-offs between output quality and computational efficiency.

My default installed Ollama came with Llama 3.2, and I chose to make it the baseline for comparison. Llama models are known for their ability to create logical, general-purpose answers on a range of tasks, and Llama 3.2 is the type of generalist that developers or companies would use without any customization much of the time. I saw it as a valuable standard against which performance could be measured for more specialized or new models like TinyLama and Deepseek.

Finally, I selected Deepseek-R1:7B as it has recently drawn attention in the AI circle and on social media for both its performance as well as availability. It had been promoted as being fast, efficient, and able to produce good quality output across tasks, and I wanted to see how this model performed in the human resource industry since it is a reasoning model. I also wished to determine whether the increasing buzz about Deepseek had any substance behind it or were largely based on marketing projections. Adding Deepseek enabled me to investigate how an extremely tuned as well as media promoted model would fare under evaluation, especially on structured, task specific applications.

All three of these models provide coverage for all scenarios: TinyLlama for resource-limited situations, Llama 3.2 as an excellent multi-tool performer, and Deepseek-R1:7B as an effective modern upstart. These model selections provide good comparisons for performance data (performance speed, memory requirements, CPU load) and qualitative aspects (quality of response, formatting, tone), ensuring both technical judgement as well as practical world validity.

Results From Exploration and Focused Experimentation

Focused Experimentation:

Prompts:

1. "Write a job description for a mid-level software engineer at a startup."
2. "Create five behavioral interview questions to assess teamwork skills."
3. "Create a Resume template for a Software Engineer candidate."
4. "An employee is consistently late and their teammate is frustrated. How should HR handle this?"
5. "List best practices for promoting diversity in hiring."
6. "Write constructive feedback for an employee who meets deadlines but struggles with team communication."
7. "Draft a welcome email for a new hire joining the marketing department, including their first-day schedule."
8. "Write a professional and empathetic message to notify an employee of a layoff due to organizational restructuring."
9. "Create a remote work policy that outlines expectations, communication guidelines, and eligibility criteria."
10. "Explain how to conduct a salary benchmarking analysis for a new role in the tech industry."

At the start of this project, I contacted a human resources consultant to help me create prompts to test an AI model's knowledge of the human resource domain. The prompts were created with practical experiences in recruitment, evaluation, and communications, and were made to cover general knowledge of the human resource domain and more in depth nuanced scenarios. The consultant only knew that there were three different models, and to rank their outputs from best to worst. When the outputs were presented, the model names were swapped to model 1, model 2, and model 3 to ensure the models were anonymous. The table below shows the rankings for all prompts.

Prompt	Ranking (best to worst)
1	Llama 3.2, TinyLlama, Deepseek
2	Llama 3.2, TinyLlama, Deepseek

3	Llama 3.2, TinyLlama, Deepseek
4	Deepseek, TinyLlama, Llama 3.2
5	Deepseek, Llama 3.2, TinyLlama
6	Llama 3.2, TinyLlama, Deepseek
7	Deepseek, Llama 3.2, TinyLlama
8	Deepseek, TinyLlama, Llama 3.2
9	TinyLlama, Llama 3.2, Deepseek
10	TinyLlama, Llama 3.2, Deepseek

Table 1.

Below is a performance table, showing model ranking comparisons through all ten prompts with three points awarded for first, two points for second, and one point for third.

Model	First	Second	Third	Total Points
TinyLlama	2	6	2	20
Llama 3.2	4	4	2	22
Deepseek	4	0	6	18

Table 2.

The best model for human resource domain questions is Llama 3.2. Llama 3.2 performed well on the prompts that covered more general human resource knowledge, but started to struggle as the prompts started to become more nuanced. However, Llama 3.2 was generally the smartest across all tasks. TinyLlama came in second, this model was able to preform at a decent level across all prompts. Deepseek was strong at prompts that required more nuance, which is due to it being a reasoning model. Deepseek generally provided good responses, but became overly verbose on some answers which made the answer lengthy.

Exploration:

Prompts:

1. General QA, What is the capital of Argentina?
2. Summarization, Summarize the following paragraph: 'Artificial intelligence is a branch of computer science that aims to create machines capable of intelligent behavior. This includes learning, reasoning, problem-solving, and language understanding.'

3. Code Generation, Write a Python function that checks if a number is prime.
4. Creative Writing, Write a short story about a robot discovering a forest for the first time.

Analysis of Strengths and Weaknesses of Each Model

To properly analyze the strengths and weaknesses of each model, I created a radar graph, table, and a group of bar graphs.

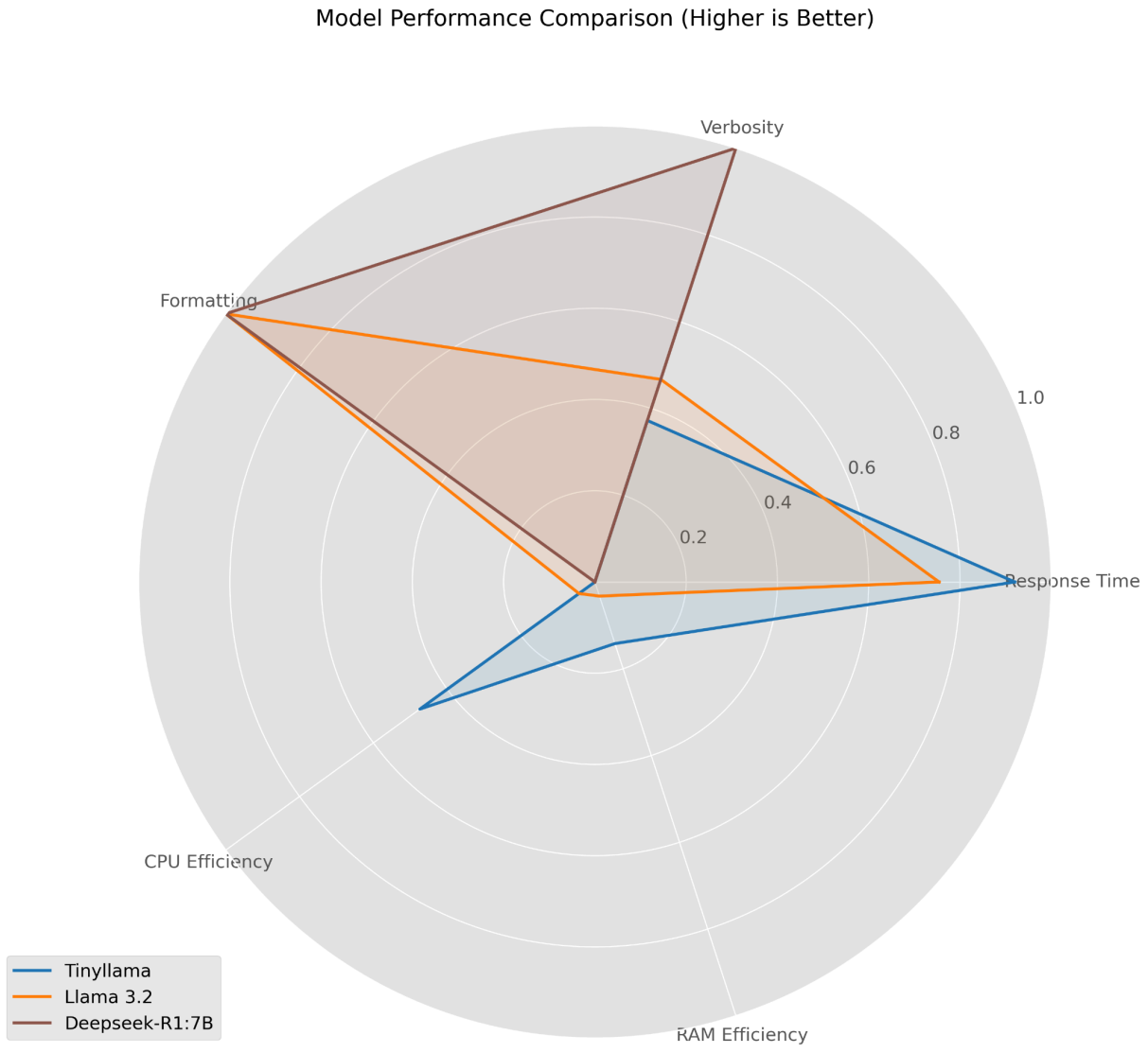

Figure 1.

Ollama Models Summary Statistics

Model	Avg Response Time (s	Avg Word Count	Uses Formatting	Overall Score	Avg CPU (%)	Avg RAM (MB)
Tinyllama	46.07	768	5	0.40	34.80	14701
Llama 3.2	11.33	359	5	0.46	33.30	14221
Deepseek-R1:7B	3.69	286	0	0.38	18.30	12612

Table 3.

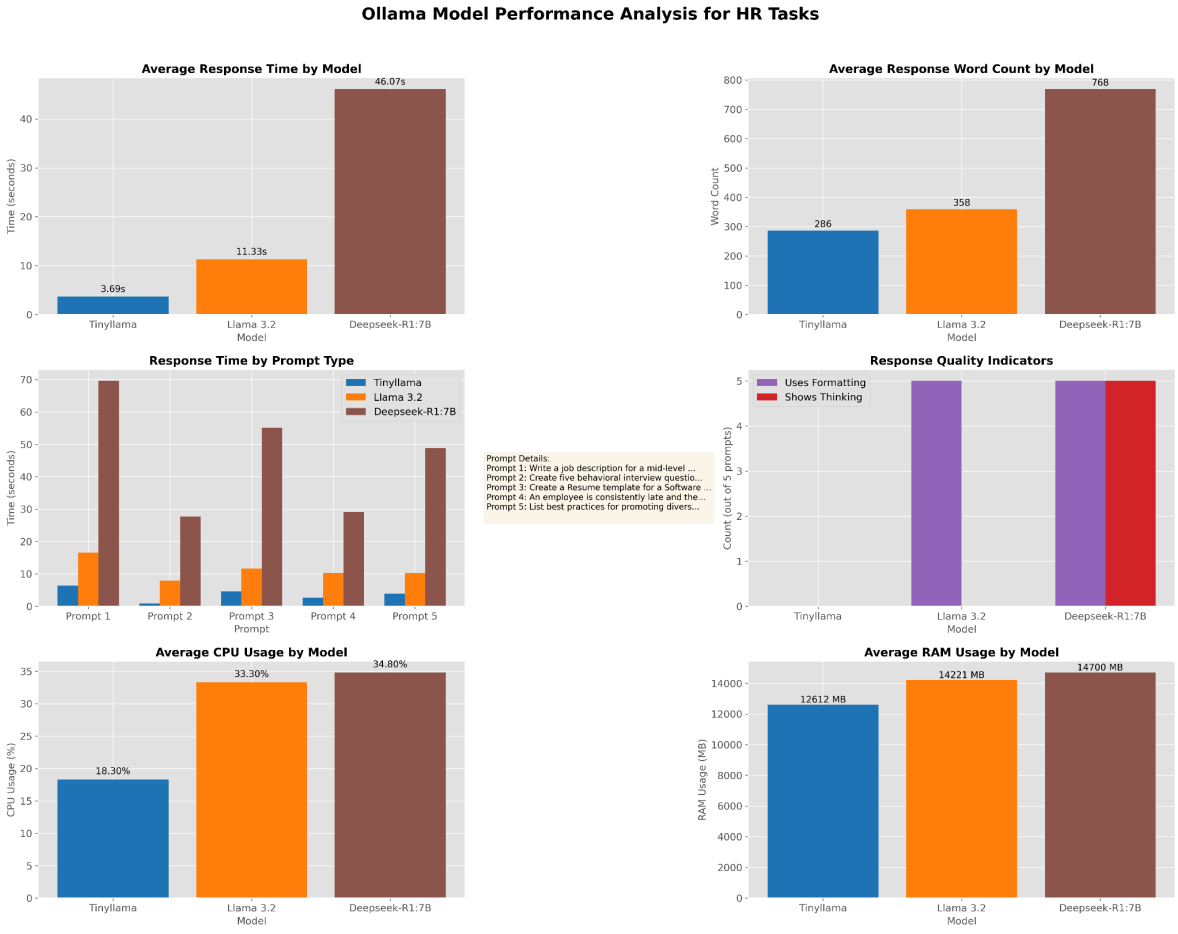

Figure 2.

Deepseek-R1:7B

Strengths: Deepseek-R1:7B is strong in providing elaborate, content rich responses, evident in having its largest average word count (768) as well as regular usage of formatting. Both this verbosity and structure are likely key in providing the clarity and completeness in qualitative rankings as it ranks among the top for more nuanced prompts. Its answers indicate having strong language generation with an emphasis on completeness as well as structure refinement.

Weaknesses: As it is so verbose, Deepseek-R1:7B is not efficient. It possesses the longest response time (46.07 seconds), the most CPU usage (34.80%), as well as most RAM usage (14,701 MB), making it the most resource-intensive model. These performance overheads can limit its usage in high latency or resource limited situations.

Llama 3.2

Strengths: Llama 3.2 is a balance between performance and quality. It ranked top overall in terms of HR rankings, with its outputs being consistently relevant, easy to understand, and properly formatted. Having the second-quickest response speed (11.33 seconds) and high CPU/RAM efficiency relative to TinyLlama, it is an effective compromise between resource consumption and output quality. It holds complete formatting and an acceptable word count, making it rank in the top 2 for most inputs consistently.

Weaknesses: Though efficient and balanced, Llama 3.2 does not overwhelmingly lead any particular metric apart from qualitative analysis. It has significantly fewer words than TinyLlama, leading us to wonder if it’s perhaps more brief, but this is either an asset or limitation based on what it’s required for. It's not as light on resource requirements as Deepseek either.

TinyLlama

Strengths: TinyLlama has the best response time (3.69 seconds), CPU consumption (18.30%), and RAM (12,612 MB) among the three models. If system efficiency and speed of performance are concerned, then Deepseek is the ideal choice. It is perfect for real-time applications or bulk use scenarios where resource consumption at minimum is essential.

Weaknesses: The reasoning comes at the expense of output quality. In the HR consultant rankings, TinyLlama repeatedly came in lowest and has the fewest words (286) and no formatting usage of any kind. Its answers feel skimpy, unorganized, or professionally unpolished, making it less adequate for applications where depth, lucidity, or polish are important. There were, however, some instances of Deepseek being overly verbose as well, resulting in responses that are lengthy or diluted.

Insights About LLM's and Their Capabilities

Working directly with several open-source large language models (LLMs) helped me better appreciate both their strength and weakness. Although LLMs can generate well-constructed, apparently intelligent answers to an incredibly broad set of tasks, they are limited as they do not think as human beings do. They're statistical machines trained on gigantic data sets, so they're good at pattern identification—but not at actual understanding.

One of the biggest surprises of this project was observing each model's different performances based on the complexity of the prompt. Llama 3.2 was best at general HR tasks and produced consistently structured, professional answers. Deepseek occasionally beat out the others on subtly complicated, logic-heavy prompts, suggesting an exceptional capacity for dealing with complicated contextual input. But this came at a cost as its answers were too lengthy, almost disproportionate, and in need of editing. TinyLlama, while swift and productive, lacked in terms of depth and structure, yet still held its own in easier tasks.

Another observation is that tone and format have a large impact on perceived output quality. Even when content was technically accurate, responses without formatting and professional tone were perceived as less polished or helpful. This was perhaps best illustrated in HR examples, in which tone and clarity can be as valued as factual correctness.

Regardless of sophistication, they do not apprehend meaning, context, or ethical implications of those tasks. They do not "get" workplace situations but instead, they provide reasonable answers based on learned associations. That shortfall leads me to further believe that LLMs should be used as assistive devices, not independent agents, in domains such as human resources, wherein empathy, discernment, and ethical thinking constitute critical components. They can facilitate workflows, inspire, and save time, but without human control, they do not represent the depth and responsibility necessary for independent decision-making.

Practical Implications and Real World Applications

This project uncovered various practical ways in which LLMs can be effectively incorporated into actual HR processes. Activities such as creating job descriptions, crafting feedback, or creating interview questions are generally time-consuming yet formulaic. Such models as Llama 3.2 were found to be capable of creating good-quality drafts for such types of activity, enabling HR professionals to concentrate on review and completion rather than creating work anew. This can greatly boost overall productivity and ease mental burden for mundane document work.

TinyLlama, while generating the least refined output, proved to have genuine potential for use in scenarios in which speed and low resource utilization are prioritized. For instance, it could be used in internal HR chatbots to provide initial answers to staff queries quickly, or in generating

templates for emails or memos. TinyLlama shows that light models can still have useful functions when outputs have to be reviewed or serve as draft documents.

Deepseek performed best in cases involving greater contextual sense or multi-step reasoning. It generated the fullest and most extensive answers, and is a strong contender for uses such as crafting empathetic layoff messages or answering workplace conflict scenarios. Deepseek has a reasoning process, where it outlines the logic and thought to get to the solution. This feature could be helpful to see how to logically approach a scenario, yet human oversight is definitely required. Its heavy resource use and verbosity would necessitate cautious integration, possibly in conjunction with additional layers for filtering to keep its outputs functional in real-world environments.

One of the strongest real-life lessons is that one size won't fit all. Rather, organizations can gain from selecting models depending on task complexity and operational requirements. For low-complexity, high-volume tasks, a smaller one such as TinyLlama can be just right. For everyday HR writing and decision support, Llama 3.2 is an excellent balance. For high-complexity, high-context communication, Deepseek can be the best choice if handled properly.

In the end, these models are instruments. They have the potential to save time, assist in creativity, and enhance consistency in communication, yet they must not be confused as substitutes for human intuition. This is especially important in an area as emotive and personal as human resources.