Large Language Model (LLM) Comparison Research

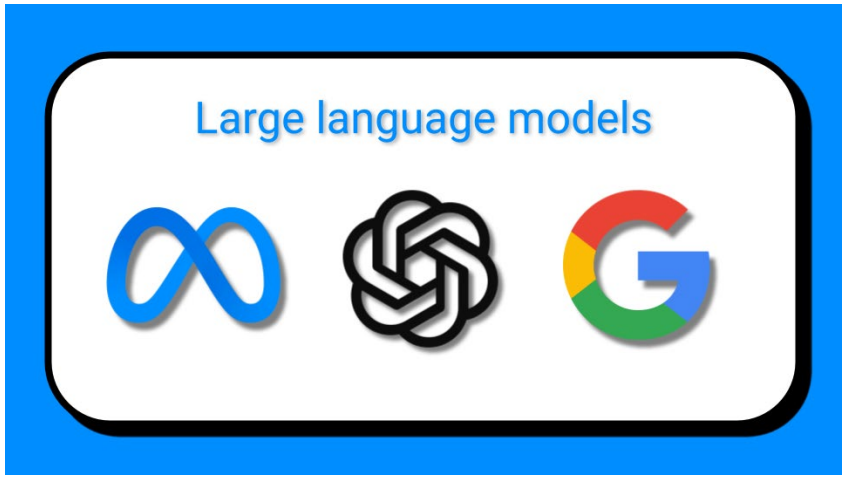
Will Fisher - Data Science Fellow

5/8/2024

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

What are the productivity savings from Large Language Models (LLMs)?

- LLMs have become much more popular and more capable over the past couple years
- Unclear how much they can help with worker productivity, and what tasks they are best for
- How much time can LLMs save workers?



Data Analysis and Executive Summaries (EXSUMs)

- Focused on data analysis tasks and executive summaries (EXSUMs)
- Performed different tasks individually, and compared time/quality to OpenAI's GPT-4
- Data Analysis:
  - Exploratory analysis
  - Modeling
  - Visualizations
- EXSUMs:
  - Comparing to existing EXSUMs

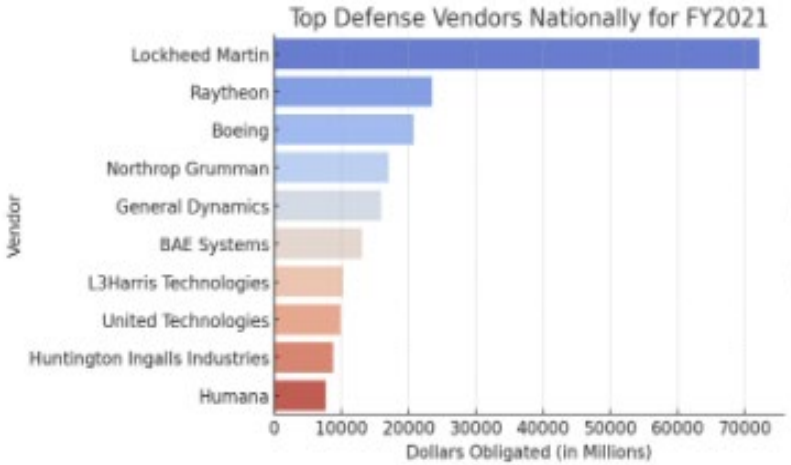
Data Analysis Tasks

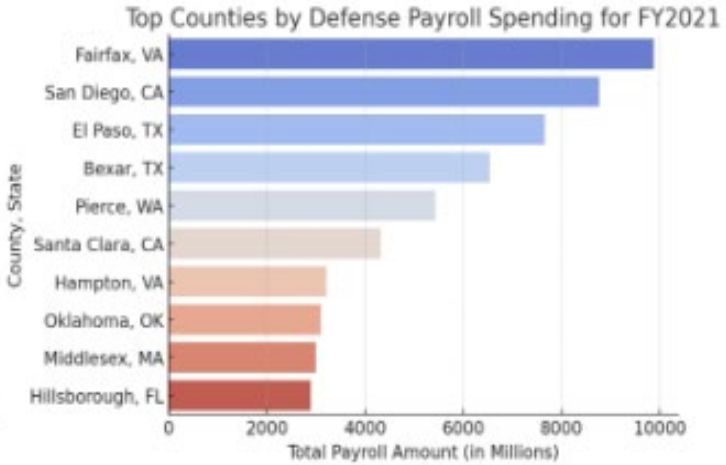
- NYC Airbnb data, Office of Local Defense Community Cooperation (OLDCC) data
- Airbnb Tasks
  - Exploratory analysis – prices, reviews, locations
  - Modeling – regression, decision tree, random forest
- OLDCC Tasks
  - Visualization – simple (bar charts) and more involved (choropleth maps)

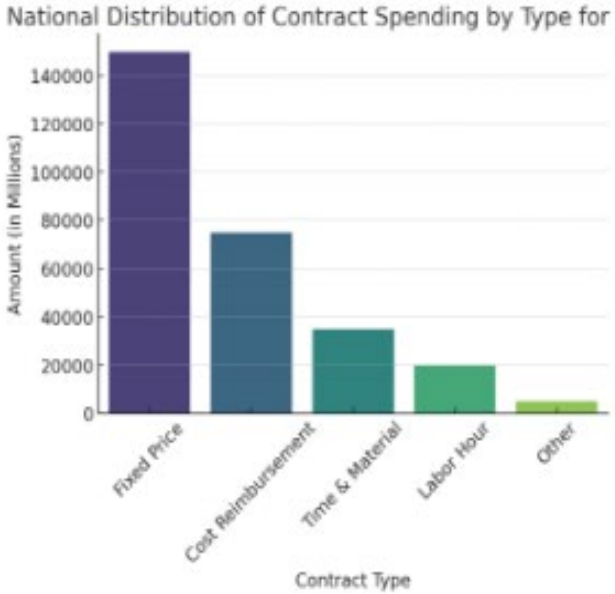
Data Analysis Results

Task	Time Saved (Minutes)	Percent of Time Saved (%)	Quality Comments
Exploratory Analysis	30	67	Thorough in investigating the nature of the different variables and made important contextual inferences
Modeling	45-60	60-80	Tested the same models that we did, but made questionable preprocessing decisions and could not run everything locally
Visualization	30-40	43-57	Quickly makes simple visualizations, but requires some human adjustments for more complex ones
Total	105-130	55-68	

GPT-4 Simple Visualizations

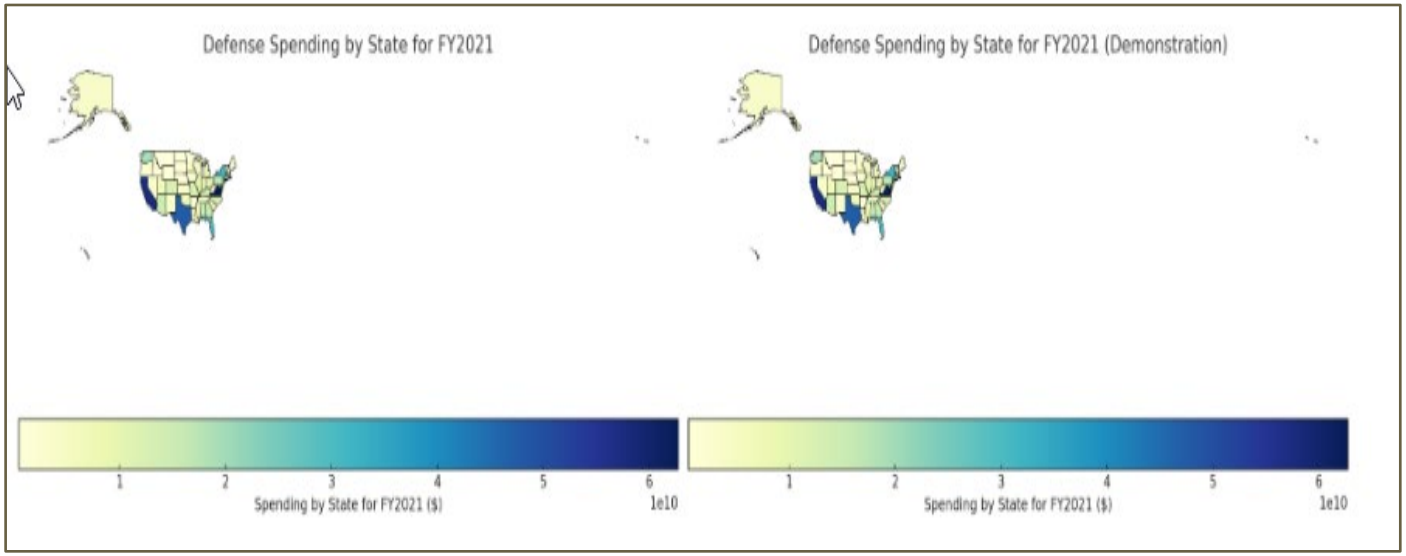




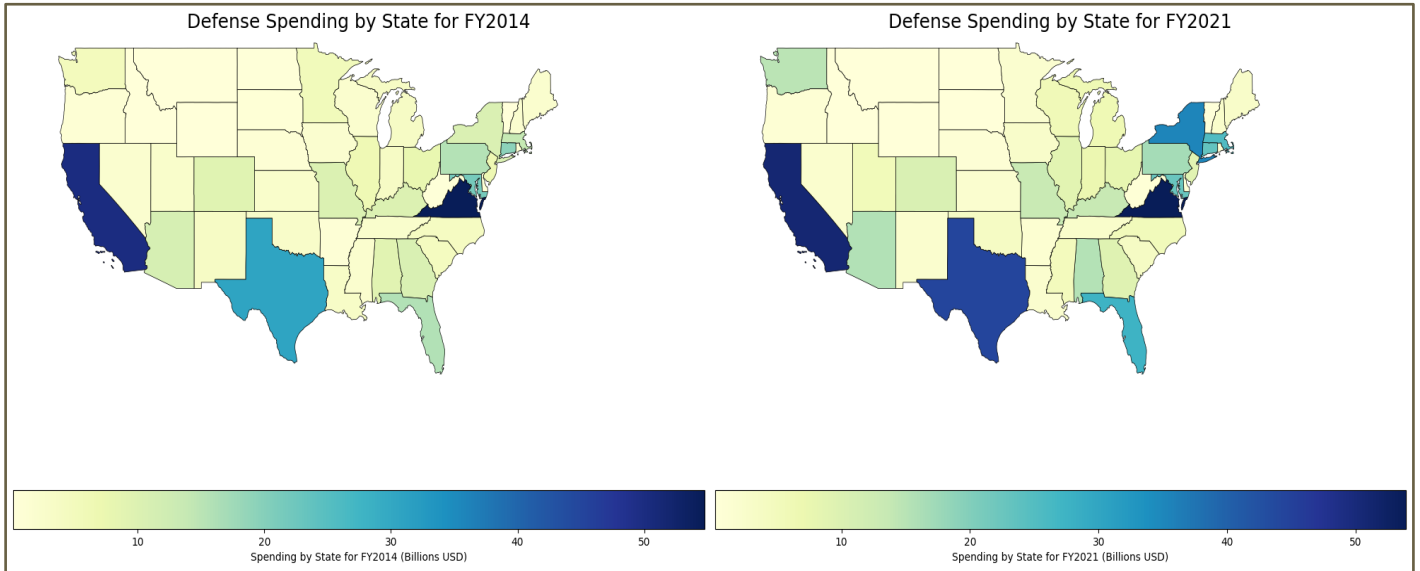


Complex Visualizations

GPT-4 Maps



Desired Maps



EXSUM Tasks

- Three IDA papers with varying levels of length and technical complexity
  - Paper 1: Factors Limiting the Speed of Software Acquisition
  - Paper 2: Forecasting Demand for Air National Guard Enlisted Initial and Technical Schooling
  - Paper 3: Quantifying and Visualizing Forecast Uncertainty with the FIFE
- Asked GPT-4 to summarize and then compared to the authors EXSUM

EXSUM Results

- Similar summary for shorter, nontechnical paper
- For the two more technical papers
  - Made up technical terms, acronyms
  - Didn't include numerical results
  - Still had solid foundation for a summary
- Reasons for more issues with technical papers
  - Encoding equations: $h^{(t)} = \underset{h \in H}{\operatorname{argmin}} E_D \left[\left(-g^{(t)}(\mathbf{x}, y) - h\left(\mathbf{x}, \boldsymbol{\phi}^{(t)}\right) \right)^2 \right]$
  - Context lengths
- Time-saving estimate: about 2 hours

Conclusions

- Data analysis
  - For simple tasks, can save most of the time spent
  - Need supervision for more complex tasks
- EXSUMs
  - Can provide base for a summary, but requires careful checks for technical papers
- Going forward
  - LLMs as assistants
  - LLMs made for specific tasks