RESEARCH SUMMARY

On-Premises Generative AI

Use of external services like OpenAI's ChatGPT can pose a significant risk to the security of sensitive and personal data. A researcher at the Institute for Defense Analyses (IDA) developed a simple new software package called OnPrem.LLM. The package easily extracts, labels, and generates text using large language models without risk of data loss. This summary explores OnPrem.LLM's features and provides an example of its use.

Not long after the release of OpenAI's ChatGPT, data security vendor Cyberhaven observed that employees across industries were sharing sensitive and privacy-protected information while interacting with the large language model. Within the defense industry, usage of such external services poses a significant security risk. However, since the release of open-source large language models like Meta's Llama, it is now possible to efficiently run ChatGPT-like models on your own machines using non-public data, without sharing information externally. OnPrem.LLM, a simple Python package developed by IDA researcher Arun Maiya using the Python programming language, makes it easier to run large language models on-premises with non-public data.

With OnPrem.LLM, you can use ChatGPT-like large language models behind corporate firewalls and within air-gapped networks with no risk of data leakage. Applications of the tool include:

- Information extraction: extract information of interest from reports.
- Auto-annotation: classify or label passages of text using only a few ground-truth examples.
- Text generation: suggest text for emails, product descriptions, social media posts, etc.
- Code generation: generate code to solve a problem given only a short instruction or description.
- Document chatting: answer questions based on content from your own documents.

OnPrem.LLM can solve each of these tasks in as little as three lines of Python code.

IDA

November 2023 | Product 3000898

```
# STEP 1: Initial Setup
from onprem import LLM
llm = LLM()

# STEP 2:  ingest the sections of the NDAA
llm.ingest('./ndaa2023')

# STEP 3:  submit a question
llm.ask('Tell me about artificial intelligence in the Coast Guard.')


# AI-Generated Answer

The Commandant is authorized to assess investment in AI innovation, test and evaluate AI
capabilities, and integrate AI into wargames, exercises, and experimentation for the transitioning
to operational use of AI and machine learning for the Coast Guard. Additionally, there will be
governance and oversight of AI and machine learning policy by the designated official who will
convene appropriate officials of the Coast Guard to integrate functional activities related to
data, AI, and machine learning. The Commandant is also required to review potential applications of
AI and digital technology for platforms, processes, and operations within two years after
enactment, identify necessary resources for improving AI use, and submit a report on their findings
to Congress.

Sources Used to Generate Answer:

SEC. 11226. ARTIFICIAL INTELLIGENCE STRATEGY.
SEC. 11227. REVIEW OF ARTIFICIAL INTELLIGENCE APPLICATIONS AND ESTABLISHMENT OF PERFORMANCE
```

In the example above, OnPrem.LLM answers questions about the 2023 National Defense Authorization Act. The software package also includes a built-in web-based user interface.

This new software offers defense industry users and others the freedom to use large language models without fear the data they use will be compromised.

Arun Maiya (amaiya@ida.org) is a research staff member in the Information Technology and Systems Division of IDA's Systems and Analyses Center. His expertise lies in natural language processing, machine learning, network science and applied artificial intelligence. He holds a doctorate in computer science from the University of Illinois at Chicago.

© 2023 Institute for Defense Analyses | 730 East Glebe Road, Alexandria, VA 22305-3086 | ida.org