GenAI In-Home PNM and Bidirectional Natural Language Interface

(NetLLM – Chat AP)

Inventors:

Zackary Foreman

Tucker Polomik

Benjamin Carlson

Casey Turtel

Field of the Disclosure

The present disclosure is a method and system that puts AI network monitoring and troubleshooting technologies in the consumer's gateway to empower the average consumer to solve their home networking issues without calling customer support and includes a user interface for chatting with the consumer.

Background

Calling the internet provider's customer service is frustrating. Users feel unheard and not helped. Customer support agents feel useless. There is a gap because support agents can't see into the home network. Customers are told to follow the same steps they already took before the call.

Currently, even the best gateway is nothing more than a water faucet that has a gauge connected to it. MSOs (Multiple System Operators) count data up and data down at the gateway and perhaps some dropped packets. But this information is not good enough to truly see the issues that a consumer faces day to day.

For power users and tech savvy users, nothing irritates them more than having less control of a device that they are required to have in their own homes. Users need a system and an app to do this configuration and/or to access a webpage for that configuration. The interfaces of the past have been cumbersome to navigate and they are full of jargon and terms that only a network professional would understand. It's easy for the tech savvy user to make a mistake that has a cascading effect, causing hours of work to debug and unwind. The internet providers' solution has been to increasingly lock down the equipment, which now requires a phone call for anything important or unusual.

The cable industry has started integrating various AI technologies into the network in order to proactively make changes for a better experience, view potential anomalies, and see security issues arise in real time. There are now tools to detect something going on at the core. However, tools that are inside the home network are lacking.

Summary

The present disclosure includes a system, tool, and way to make users' home networks easy to interact with. No more complicated GUI's, no more apps, customer phone calls, etc. Note that the present disclosure includes multiple embodiments and inventions referred to herein as the "disclosure," "method," "tool," and "system," all of which can be used interchangeably. In some embodiments, the user can use the present system to do everything they want through text and voice. Some embodiments of the present disclosure are like a personal customer service agent in a box.

The primary components of this system and method include removing the need to call the internet provider's customer service when an internet issue arises, as well as creating a more intelligent gateway onto the MSO's network. In essence, the present disclosure uses network communications and sends them into a trained large language model ("LLM") (e.g., a GPT engine) and then the trained LLM diagnoses the user's network, connection, website the user is using, and/or server the user to which the user is connecting over the network, among other things.

In one embodiment, operators (e.g., MSOs) deploy an AI-based home network assistant that acts both proactively and reactively to problems in a subscriber's home network. The assistant would also have a natural language interface to be able to have bidirectional conversations (voice, text, etc.) to users in the home.

One embodiment of the present disclosure is a tool at the user's home that even a novice, non-tech-savvy user can use. These in-home tools can run and give a technical output. A trained LLM can be used to translate these complex technical outputs to information even technologically illiterate users can understand.

Embodiments of the present disclosure include a chat system that can, among other things: (1) convey that a website is down and why it isn't working; (2) determine why the internet is not working on the user's device; (3) update the Wi-Fi password; and/or (4) create a guest network, e.g., for short-term guests or the plumbers who are currently doing work at the user's home and they need Wi-Fi so they can bill correctly.

At a high level, the present disclosure is a tool/system that turns on the user's wireless access point and acts as an interface to their system, collects metrics (packet captures), sends data up to a server, engineers a prompt—along with other information—into a request for a specifically trained LLM for diagnosis, diagnoses the problem, then informs the user of the diagnosis and/or informs the MSO of any network issue. If the problem is on the access point ("AP"), then the present tool/system works on the problem to fix it. If the problem is outside of the local area, then the tool/system will inform the MSO and user.

The tool/system (sometimes called the "assistant" herein) learns network usage patterns in the user's home and optimizes things proactively. The tool/system would have a binary classifier that determines whether or not a networking anomaly is taking place inside the home or somewhere in the operator's network at large. If the problem is within the user's home, the tool/system will proactively give its best effort for fixing the problem or make suggestions to the end user, for example, "I see the Wi-Fi signal to your MacBook is weak, try moving closer to your access point."

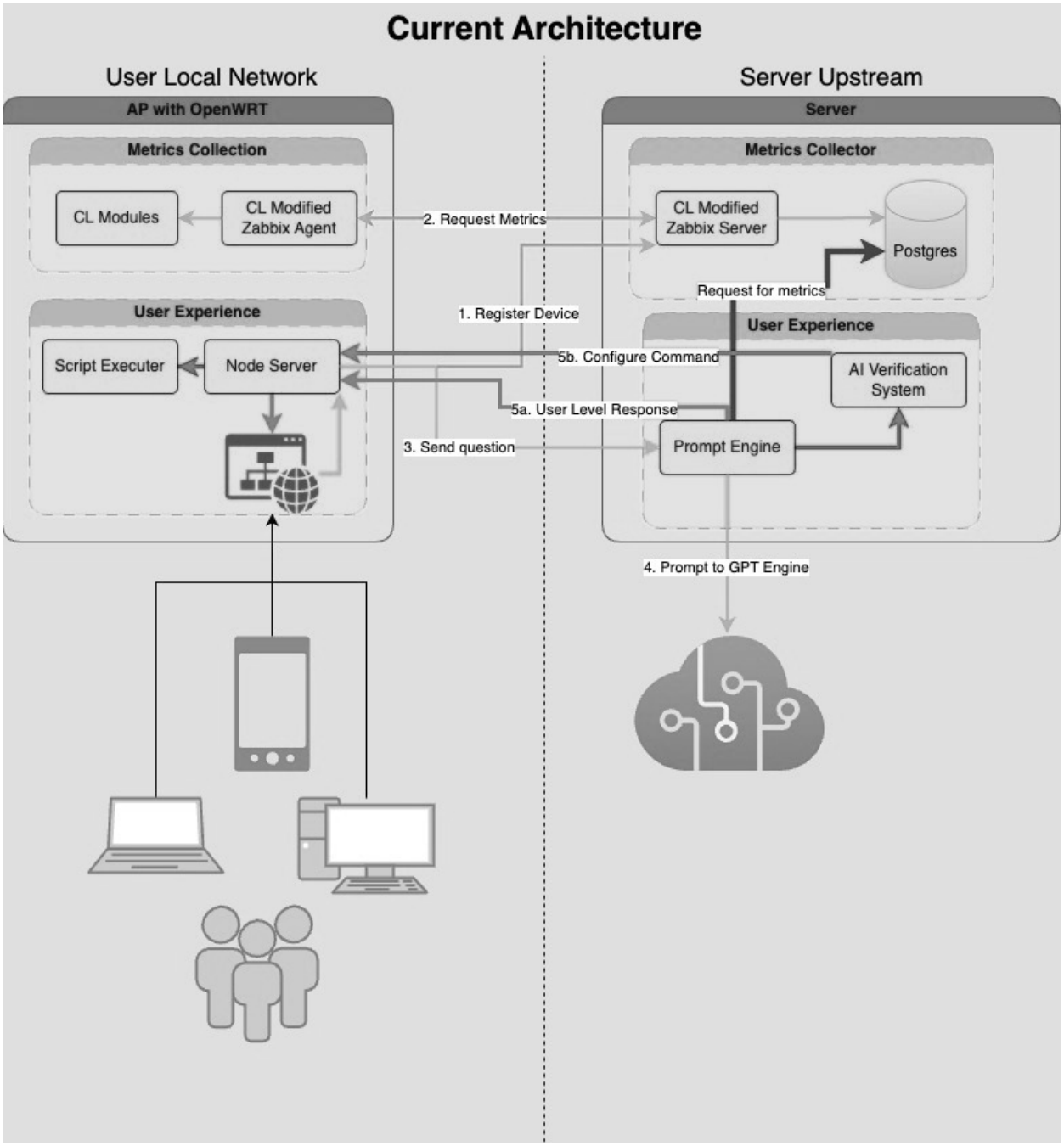
Since users can also speak to the tool/system, they can say things like, "I have an important Zoom meeting at 3PM, can you prioritize my service?", whereby the tool/system will place the user's laptop on a higher bandwidth Wi-Fi channel and limit throughput for all other devices in the home for the time window of the Zoom meeting.

This would give end-users a sense of being "in the loop" in terms of network troubleshooting. Subscribers mostly feel left in the dark, unsure why their network is acting up, causing resentment of their service providers. This tool/system includes a chatbot / app / user interface that puts a face to the service provider and totally changes the subscriber-provider paradigm from one of exploitation and resentment to one of cooperation and perceived teamwork.

Advantages of embodiments of the present disclosure are that MSOs will have less truck-rolls, less customer care contact, and higher subscriber retention.

Detailed Description

Currently the system/method of the present disclosure has an architecture that looks like the below figure.



This will all be able to run inside a user's home, on something similar to a Nvidia Jetson. It seems complicated, but the main take away is that there is a website hosted on the home gateway / AP. This is then accessible from any device capable of utilizing a web

browser, e.g., a phone, tablet, laptop, desktop, etc. When the user logs onto the site or opens the app, this kicks off a series of background events. Registering that current device along with fetching metrics on that device: RSSI, PCAPs (packet captures), etc. This information is then stored for further analytics.

From the website, the user types in their problem or request. The system then passes this message in the prompt engine. The prompt engine determines whether it is a configuration or a diagnosis problem, then gathers information that is applicable to that request: current configuration state and/or data metrics. From this, the prompt engine crafts the message to the trained LLM.

The message to the user is then sent back to the user, with any possible configuration changes first being sent to the system's verification system, which concurrently checks for semantic correctness. However, in some embodiments, a virtual replica of the in-home AP can be created. Once verification is performed, this is then sent back to the home AP for the configuration value to be updated, asking the user if this has solved their issue(s).

Embodiments of the present disclosure can:

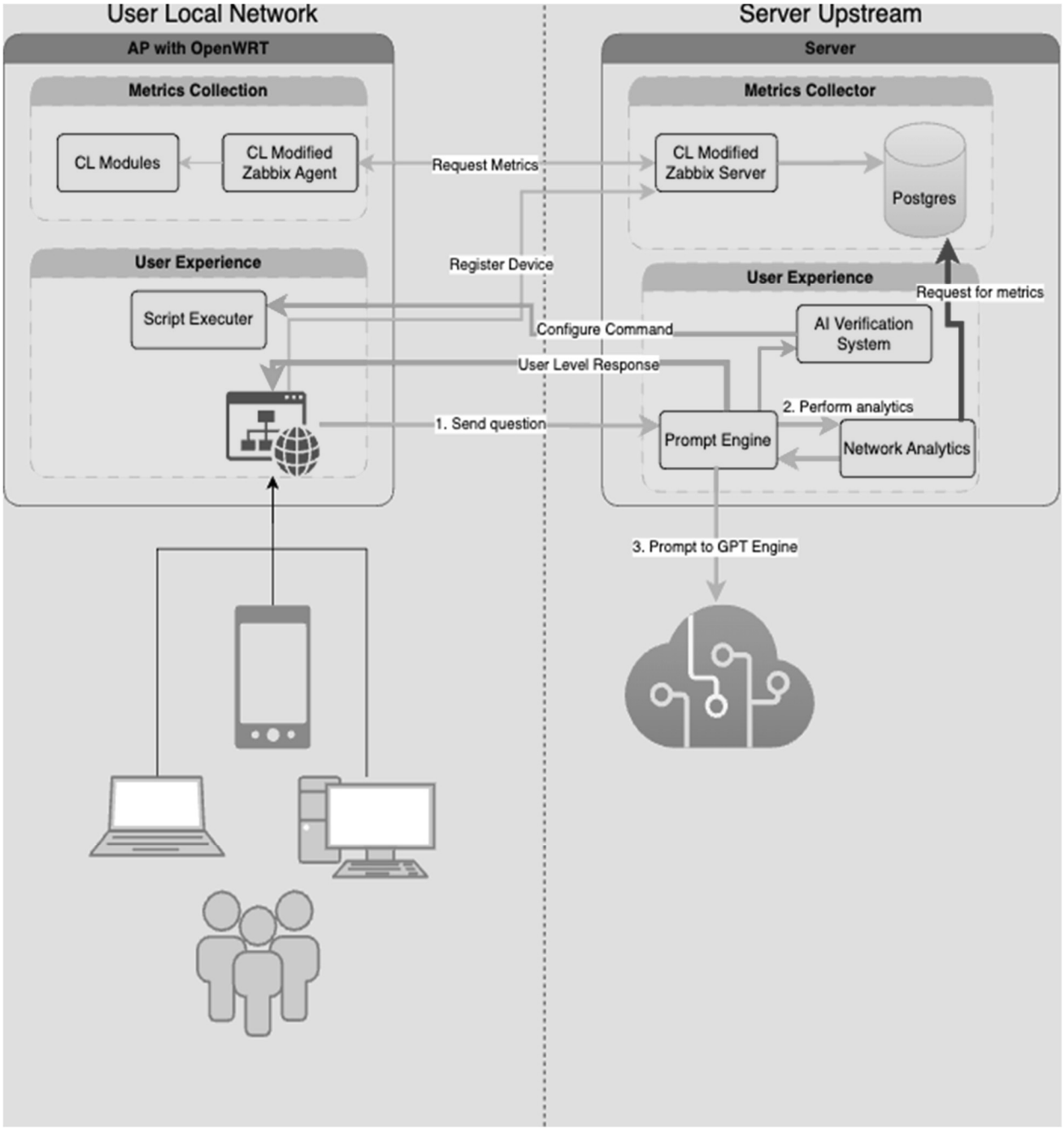
1. Determine poor signal strength: The system collects the metrics to determine this and a new process is added into the server portion. The LLM (e.g., GPT engine) cannot solve every problem. Therefore, another layer of AI is added to do analytics that the system can then feed into the trained LLM to translate back to the user what is the issue. This plays into one of the initial four cornerstones of the goal to add other AI proactive network monitoring (PNM) processes into the home network.

2. Diagnose a slow internet experience: While browsing, embodiments of the present disclosure can tell the user the website is responding slowly and it's nothing that the MSO or user can do, i.e., there is something wrong with the website and/or the website's server.

3. Diagnose a broken backhaul: in other words, maintenance on the MSO system needs to happen. Embodiments of the present disclosure also let the users know that the problem is a monumental task. With the present tool/system, it can notify the user, when

they ask, that scheduled maintenance is being performed or there is a problem with the backhaul, but someone at the MSO has been notified.

Here is an image of how the tool/system has evolved to accomplish the above steps:

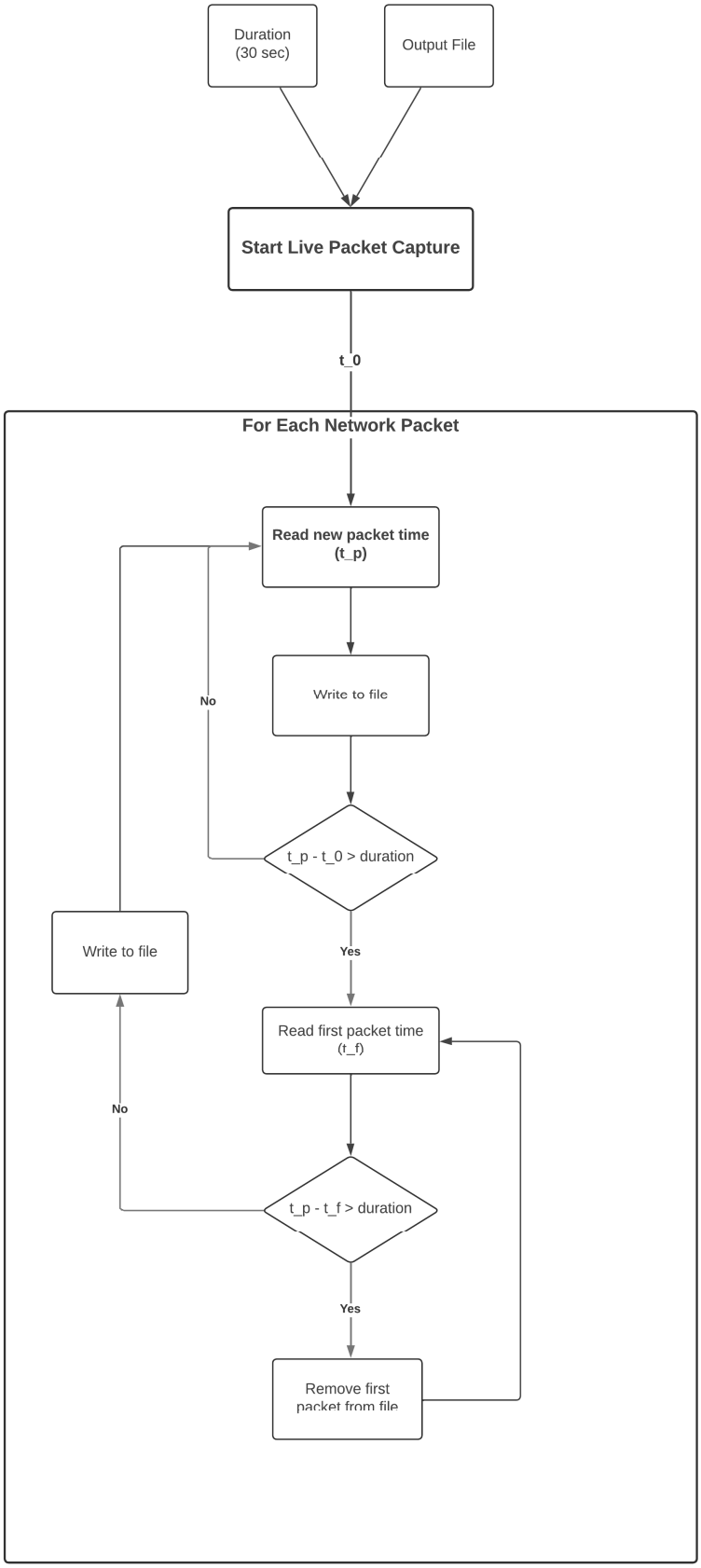


The present tool/system will create a new process on the server that is a network brains and analytics portion. The same workflow would apply: the user sends the request

and then the request is intercepted, modified with relevant information, processed, and returned.

Using a specifically trained LLM the present system uses the trained LLM to read PCAP traces and confidently come to a diagnosis. Current metrics that are collected by default are not enough to debug what is going on in a network and they are not at a high enough frequency nor a high enough fidelity to solve the problems users face that lead to negative experiences. Therefore, embodiments of the present disclosure include higher fidelity and frequency of metrics analyzed and collected. The PCAP data and context of access points are fed into the specifically trained LLM for diagnosis.

One embodiment of the rolling PCAP collector method is below.


Duration
(30 sec)
Output File
Start Live Packet Capture
t_0
For Each Network Packet
Read new packet time
(t_p)
Write to file
No
t_p - t_0 > duration
Write to file
Yes
Read first packet time
(t_f)
No
t_p - t_f > duration
Yes
Remove first
packet from file

Embodiments of the present disclosure can also be used for new network health analytics along with security features.

Embodiments of the present disclosure include a NetworkGPT or NetworkLLM (aka "NetLLM"), i.e., a network centric LLM or GPT engine that can analyze network traffic to find anomalies and mistakes in the communication. A GPT engine can read PCAPs and diagnose; however, the present disclosure takes it a step further by feeding the LLM network language and training the LLM on network language. Embodiments of the present disclosure include machine learning (ML) models specifically created for this tool/system and the ML models are used for classification and configuration versus diagnosis questions. Thus, a unique multipurpose tool has been created that can find the errors in the network traffic that is occurring. Identifying and pinpointing errors in PCAPs is a long and tedious process.

Some embodiments of the present disclosure also have a tool that can generate very real traffic. Instead of needing to be tested with complicated and expensive tools, the present disclosure has the NetworkLLM (i.e., the "assistant") do it for the system.

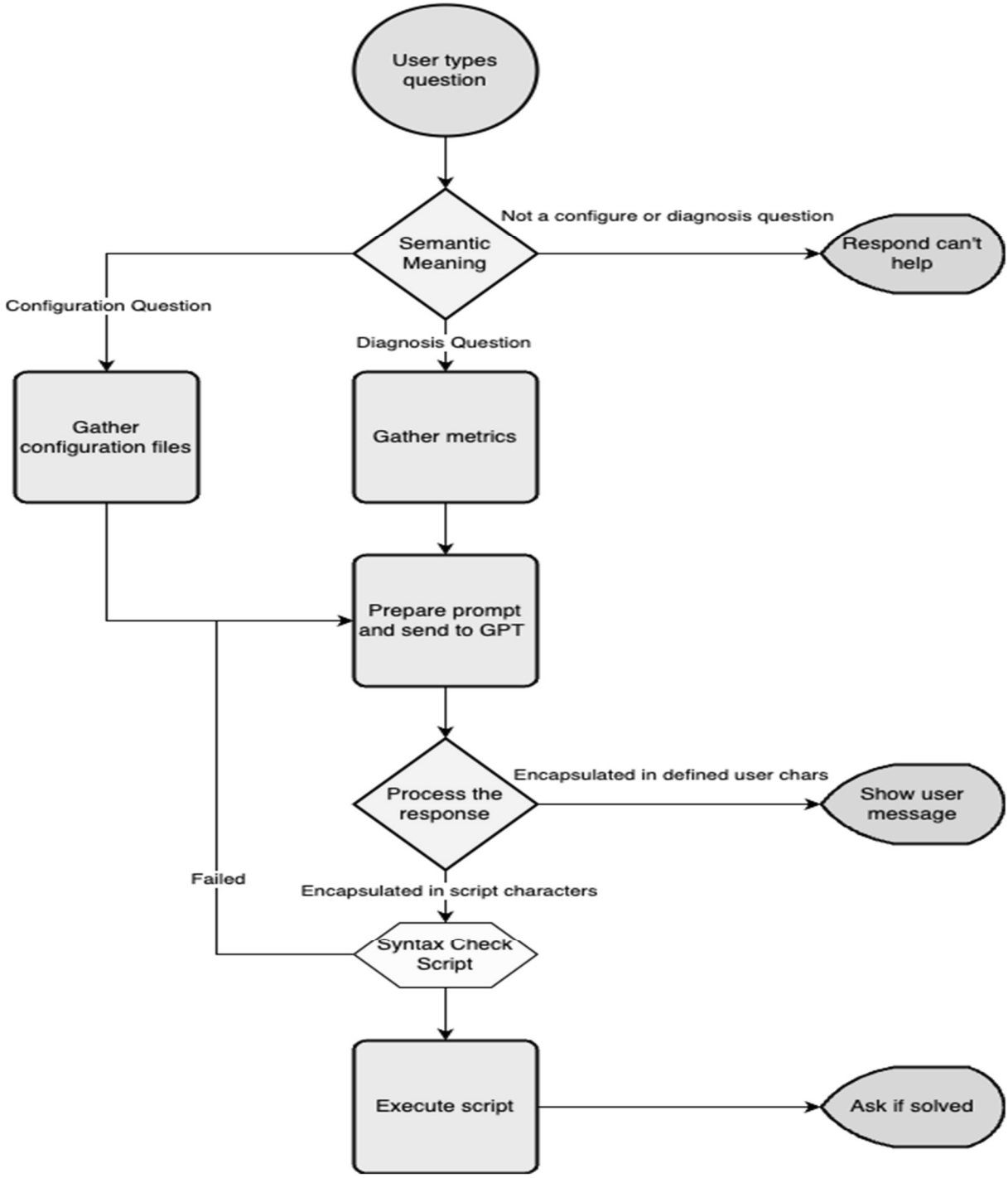
Some embodiments of the present invention are shrunk such that they can run on a Nvidia Jetson. This means that the present system could be an all in one device: gateway, AP, smarts/brains. Special purpose devices can run locally in the network, removing the need for a server in the cloud.

In some embodiments, the present disclosure is an in-home proactive network: instead of waiting for the user to say something is wrong, the system can identify a potential issue and apply a solution before the user is aware. For instance, modifying the traffic such that a video conference (e.g., Zoom) session stops buffering.

Additionally, embodiments of the present disclosure can include NaaS (network as a service) integration. Third parties can ask the system for information without needing to enter a user's home device. Thus, privacy can be applied by preserving techniques before it ever reaches these third-party individuals. Also, the tool/system can implement custom security requests for individual users, which creates an easy interface to then request that information to be applied for that user, for example, the user is using public Wi-Fi at a

coffee shop. Specifically, third party vendors could pay to access the information and analysis to build better products as well as help their customers using MSO's networks.

One method according to the present disclosure is as follows:



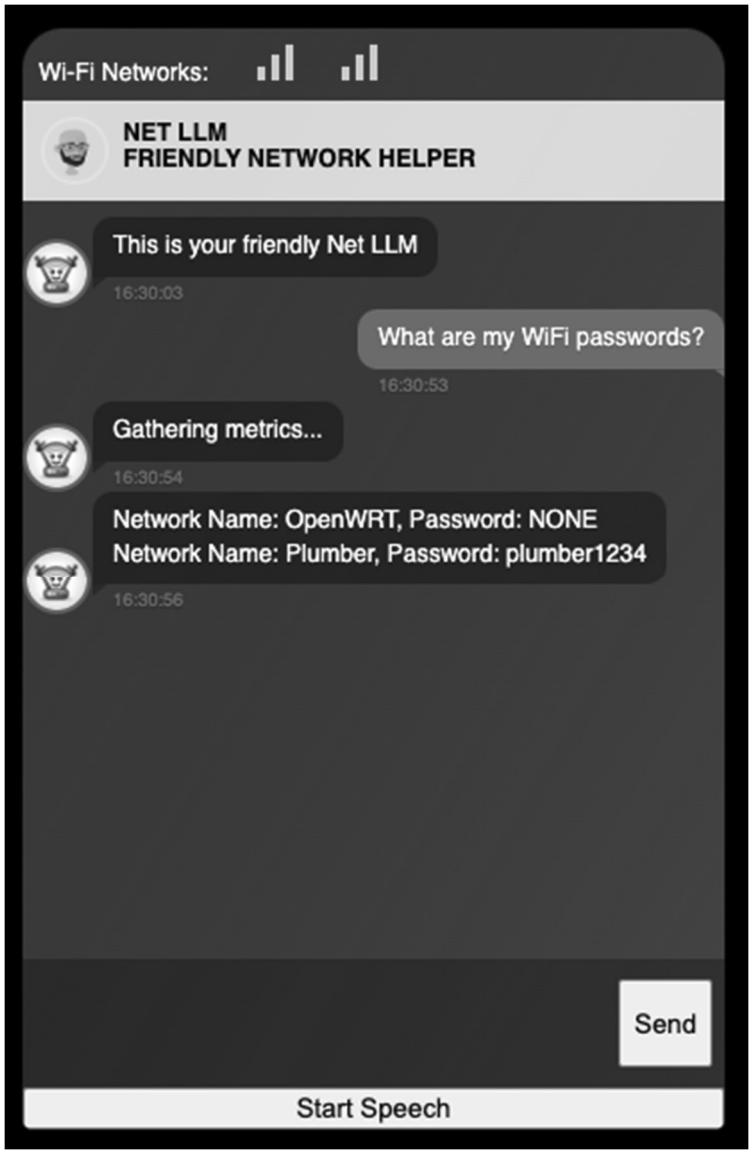
In the above method, the user's question is identified as either a configuration question or a diagnosis question or is marked as a question that the chatbot/app cannot

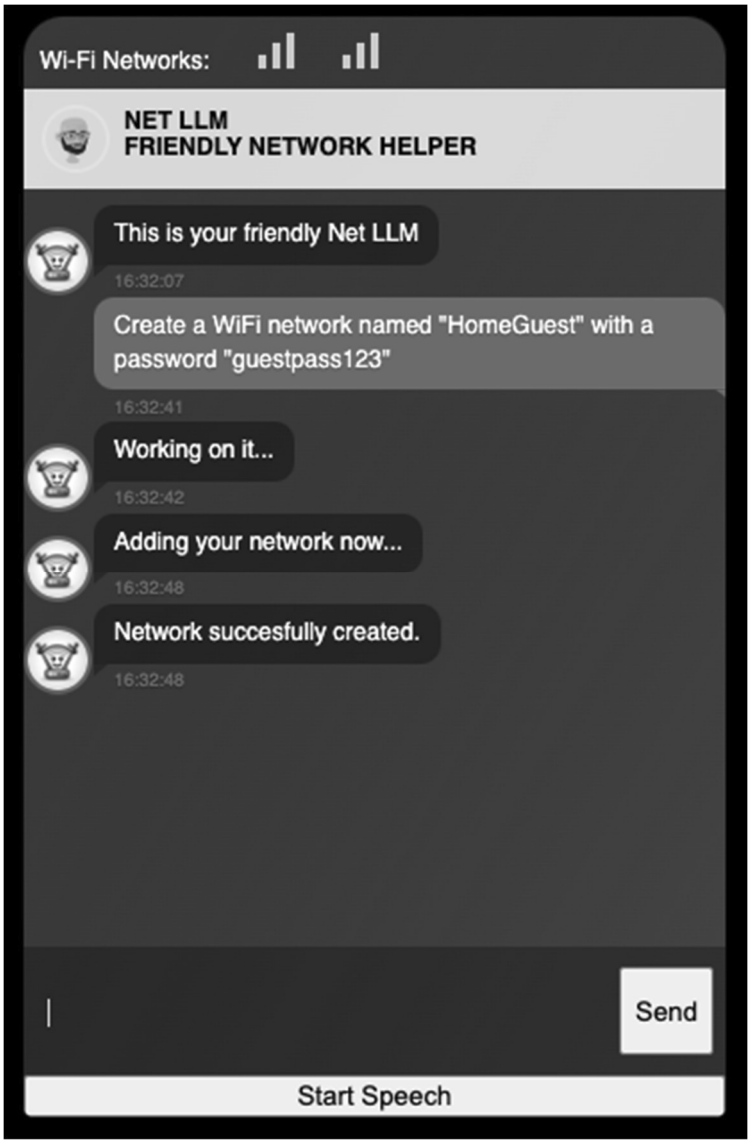
answer. After this identification, then the system or method decides if a package capture needs to occur and the details around the package capture.

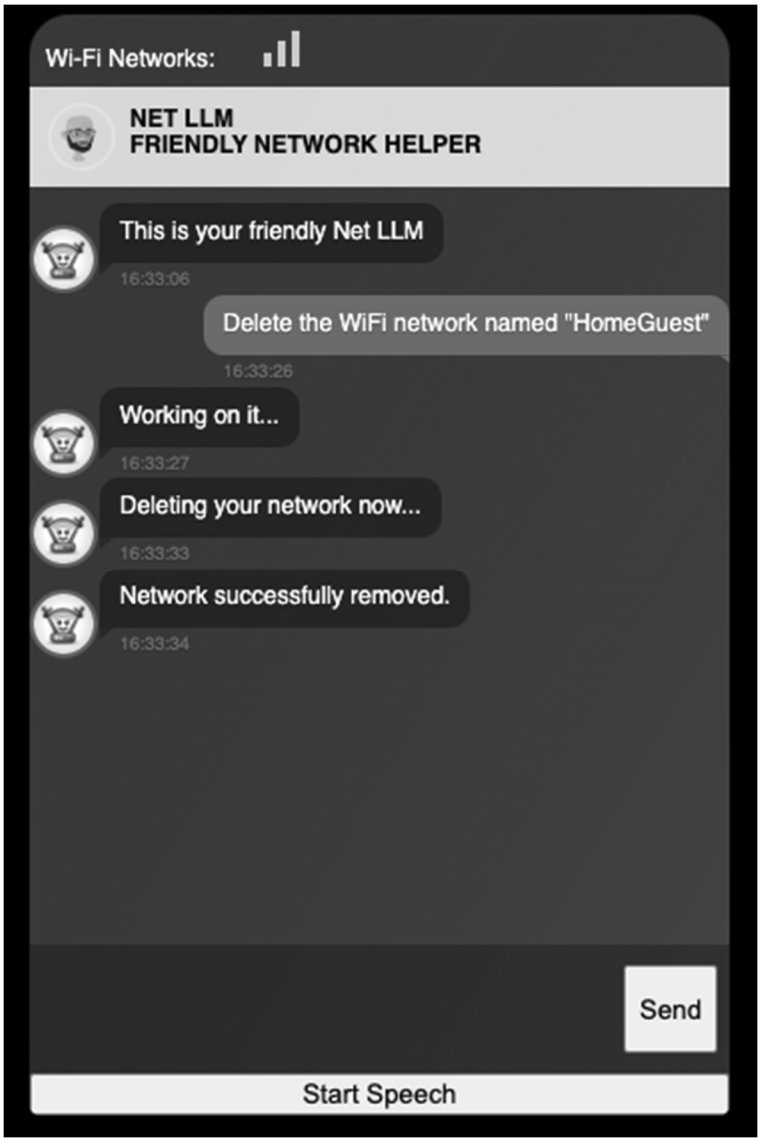
For many MSOs, the biggest concern with LLMs and GPT engines is repeatability and prevention of user manipulation. The present method handles this potential problem by not allowing any conversation that is not related to debugging or configuration changes in order to prevent the users from manipulating the chat agents to agreeing to free products or services. For handling repeatability, the present system removes ambiguity in the request as much as possible. In other words, it is more of a fill in the blank scenario, rather than a creation of an entire process from scratch. Accordingly, the present system is able to get repeatable responses utilizing a specifically trained LLM, i.e., the NetLLM.

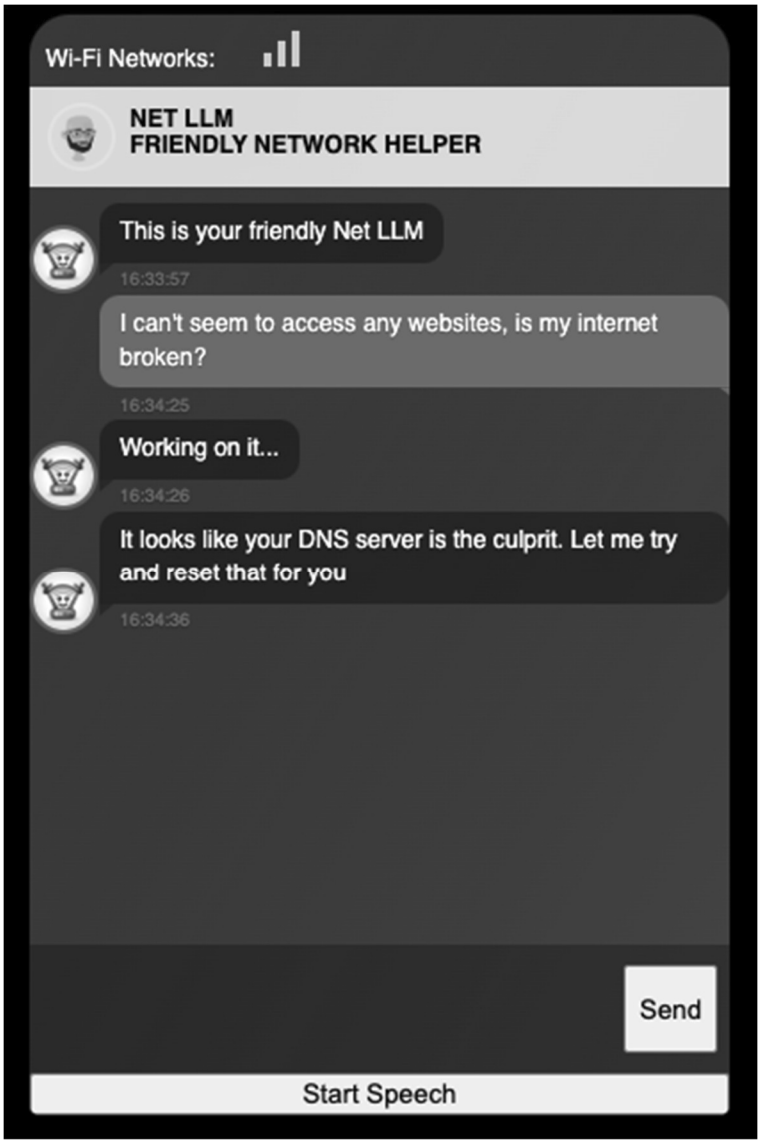
Another part of the present disclosure includes getting reliable scripts for the chatbot / chat interface with which the user interacts. The specifically trained LLM includes semantic meaning modules to decipher the user's questions and also provide a response to the user in plain English.

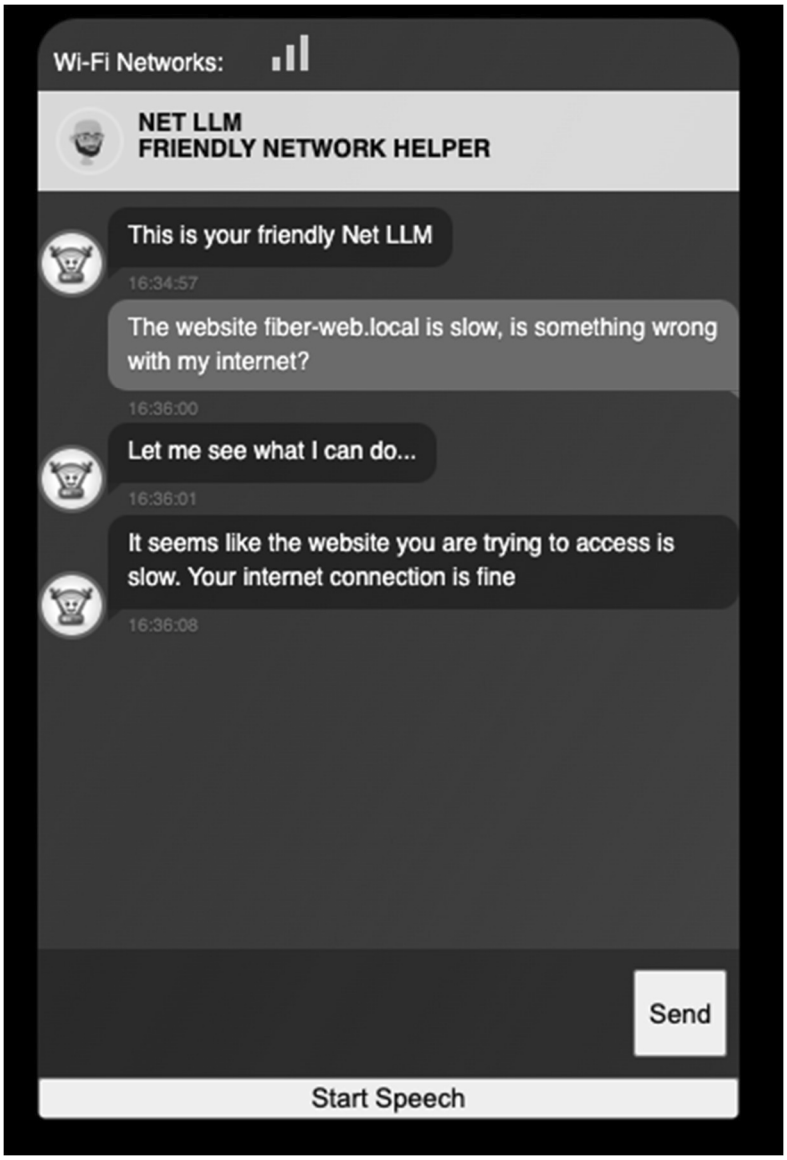
Example NetLLM interface

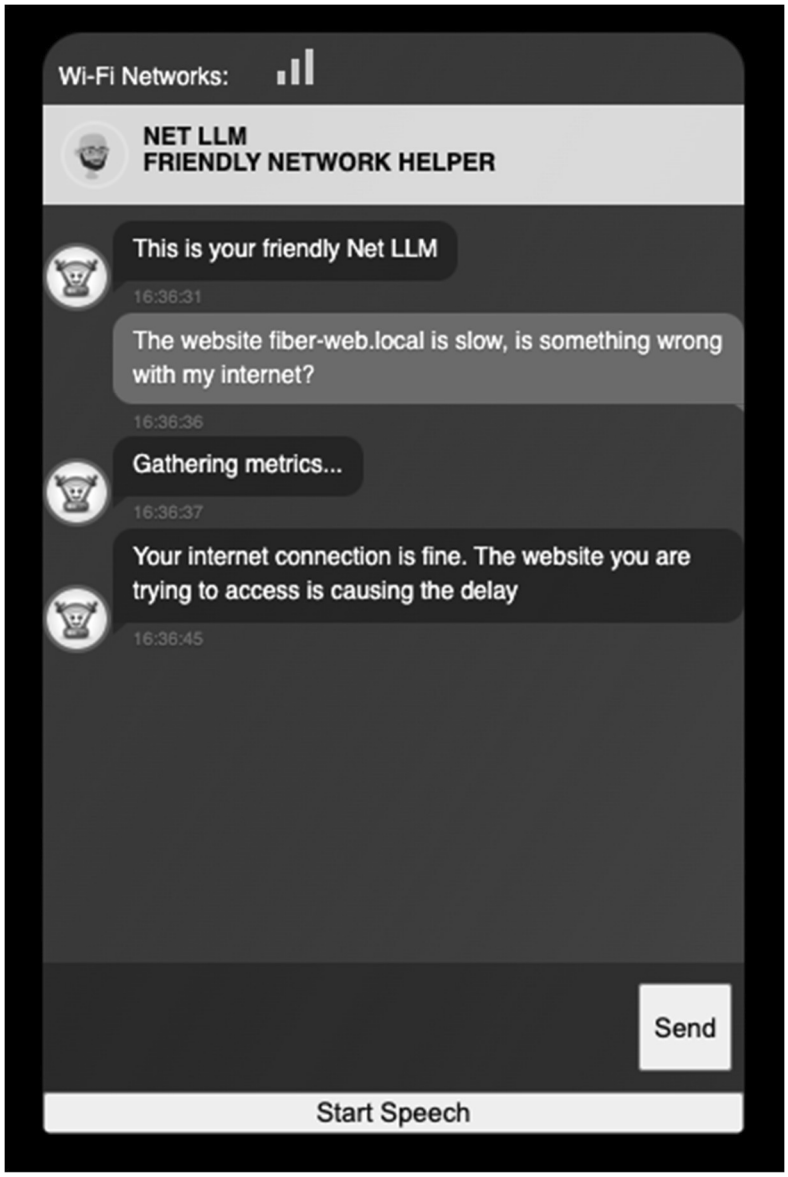



Wi-Fi Networks:
NET LLM
FRIENDLY NETWORK HELPER
This is your friendly Net LLM
16:32:07
Create a WiFi network named "HomeGuest" with a password "guestpass123"
16:32:41
Working on it...
16:32:42
Adding your network now...
16:32:48
Network succesfully created.
16:32:48
Send
Start Speech


Wi-Fi Networks:
NET LLM
FRIENDLY NETWORK HELPER
This is your friendly Net LLM
16:33:06
Delete the WiFi network named "HomeGuest"
16:33:26
Working on it...
16:33:27
Deleting your network now...
16:33:33
Network successfully removed.
16:33:34
Send
Start Speech


Wi-Fi Networks:
NET LLM
FRIENDLY NETWORK HELPER
This is your friendly Net LLM
16:33:57
I can't seem to access any websites, is my internet broken?
16:34:25
Working on it...
16:34:26
It looks like your DNS server is the culprit. Let me try and reset that for you
16:34:36
Send
Start Speech


Wi-Fi Networks:
NET LLM
FRIENDLY NETWORK HELPER
This is your friendly Net LLM
16:34:57
The website fiber-web.local is slow, is something wrong with my internet?
16:36:00
Let me see what I can do...
16:36:01
It seems like the website you are trying to access is slow. Your internet connection is fine
16:36:08
Send
Start Speech


Wi-Fi Networks:
NET LLM
FRIENDLY NETWORK HELPER
This is your friendly Net LLM
16:36:31
The website fiber-web.local is slow, is something wrong with my internet?
16:36:36
Gathering metrics...
16:36:37
Your internet connection is fine. The website you are trying to access is causing the delay
16:36:45
Send
Start Speech