

Estimating API calls to LLMs via Supervised Learning on Packet Captures

Zephaniah Roe

October 2025

Project title

Estimating API calls to LLMs via Supervised Learning on Packet Captures

Project Participants

It is only me:

1. Name: Zephaniah Roe
2. CNET: zroe
3. Project role: everything

Project summary

When I use an open source GitHub repo that makes calls to LLMs, I put in my API keys and hope for the best. I let the code run for a while, and the package may or may not give me statistics of how many API calls I have made, how many tokens I have received and how much estimated money this has all costed me so far.

One solution is to just check how many tokens I have spent in the UI of the service provider. The issue is that I may have multiple accounts, and different API keys from different providers so this is hard to check.

Another solution is to go into the package, figure out what is going on and add print statements for the information I want. This task can actually be extremely difficult.

The above solution also isn't strictly necessary. I don't need exact details of how many tokens I have received, I just want a rough estimate to give me piece of

mind that my code is still running and that I haven't spent all my API credits on it.

I proposed training an ML model to generate this information for me based on a packet trace.

Data

I can run a few controlled packet captures of me running an LLM (and I can document when I receive my tokens and how many I receive in my script).

Machine learning

I will start with an MLP. This worked well for the packet capture assignment and I think it is a good place to start. I would also be interested comparing with a simple regression model or other more complex neural networks.

Evaluation

The most basic thing I would like is the number of tokens based on an API response. I would also like to see if I can predict total cost based on the packet received (this is non-trivial because the packet will be encrypted and getting the features correct will be essential to pulling this off).

The estimated cost and tokens can be framed as a regression task but I will likely make it into a classification task where each class is a range of values. This should make the model more reliable.

Deliverable

A model that can take in the packet trace and provide total estimated tokens and cost. The model will be called multiple times (one time for each flow / API exchange).

Learning objective

I would like to learn how apply what I have learned in class to a problem in my real life. I also think there is a lot to learn from investigating a problem where I don't know much about what to do beforehand. The video resolution project was easier because it was properly scoped.