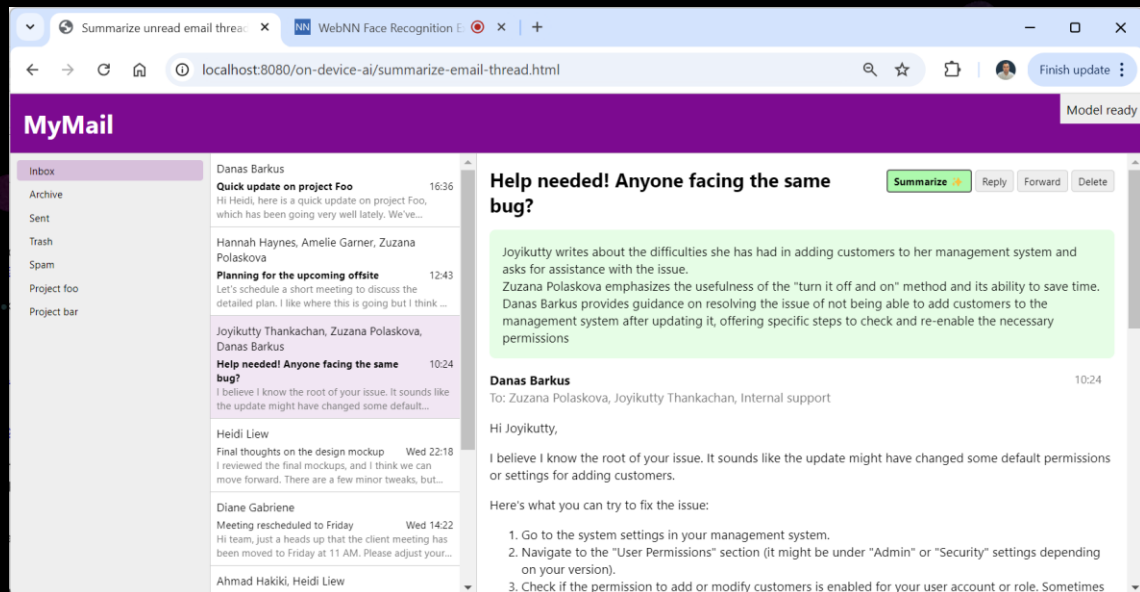


AI on the Web Platform

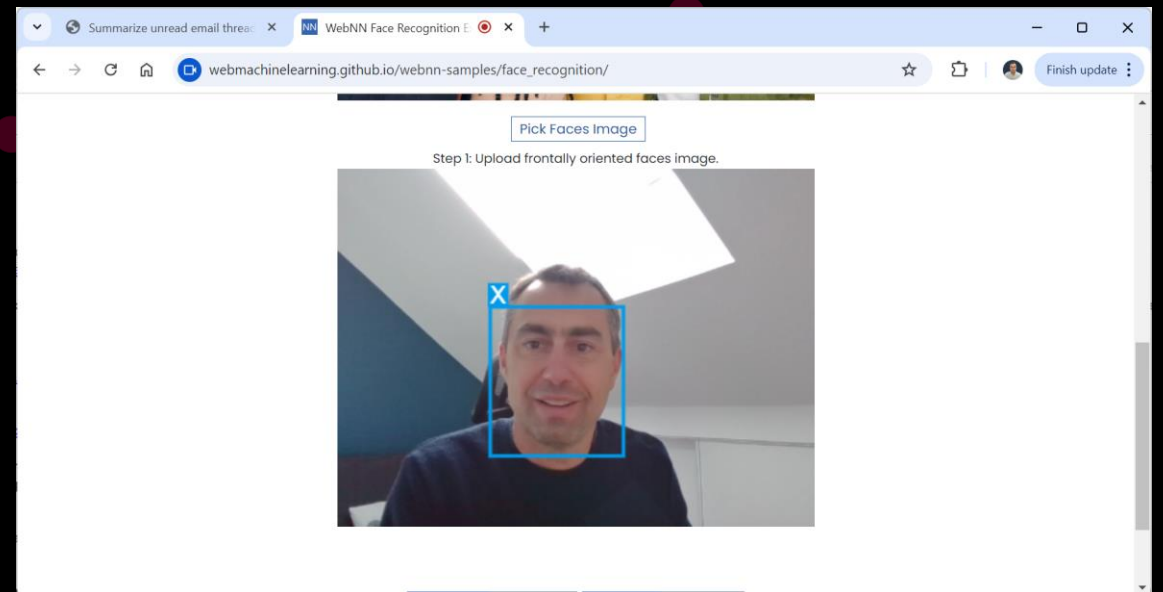
Early explorations with on-device AI

Patrick Brosset, Smashing Conference, October 2024

Demo time!







The screenshot shows a web browser window with the address bar at `localhost:8080/on-device-ai/summarize-email-thread.html`. The page title is "MyMail" and it includes a "Model ready" indicator. On the left is a sidebar with folders: Inbox, Archive, Sent, Trash, Spam, Project foo, and Project bar. The main content area displays an email thread. The selected email is from Danas Barkus with the subject "Help needed! Anyone facing the same bug?". A green summary box is overlaid on the email content, containing the text: "Joyikutty writes about the difficulties she has had in adding customers to her management system and asks for assistance with the issue. Zuzana Polaskova emphasizes the usefulness of the 'turn it off and on' method and its ability to save time. Danas Barkus provides guidance on resolving the issue of not being able to add customers to the management system after updating it, offering specific steps to check and re-enable the necessary permissions". Below the summary, the email body text is visible, including a response from Danas Barkus and Heidi Liew. At the top right of the email content, there are buttons for "Summarize", "Reply", "Forward", and "Delete".



The screenshot shows a web browser window with the address bar at `webmachinelearning.github.io/webnn-samples/face_recognition/`. The page title is "Pick Faces Image". Below the title, there is a text instruction: "Step 1: Upload frontally oriented faces image." Below this instruction is a video frame showing a man's face. A blue bounding box is drawn around the man's face, and a small blue 'X' icon is positioned at the top-left corner of the bounding box. The video frame is part of a larger interface, likely for demonstrating face recognition capabilities.

What if?

- What if this could run on users' devices?
 - No privacy concerns 
 - Lower latency 
 - No cloud AI cost 
 - Offline support 

What if?

- What if webpages could use an API to do this directly?
 - No framework to use
 - No model to download
 - Simple usage

Wait, let's take a step back

- The web and AI

Wait, let's take a step back

- The web and AI
- The web platform and AI?

It's already happening

"Bring Your Own Model" - BYOM

- WebGL
- WebGPU
- WebNN

- TensorFlow Lite
- ONNX Runtime Web

It's already happening

Google's "built-in" Prompt() API

- Browser-provided model
- Simple API
- For experimentation only ⚠

Our interest

- We want the web to succeed
- Devs want the web to be at par with native
- We want to understand user's needs first
- Not just Language Models
- We want to make it work with the web

You can help 🙏

- Ideas? Feedback? Use cases? Thoughts?
- Come find me during the conference
- Message me:
 - Mastodon: @patrickbrosset@mas.to
 - Email: patrick.brosset@microsoft.com
- Anonymous form: aka.ms/web-ai-feedback

Microsoft Edge





Come chat about:

- On-device AI for the web
Ideas/use cases? aka.ms/web-ai-feedback
- Progressive Web Apps and PWA Builder
- Web Platform stuff




On-device AI for the web

What if you could run AI models on a web page?

 Privacy-preserving
 No cloud AI cost

 Low latency
 Offline support

 Ideas/use cases?
aka.ms/web-ai-feedback

You already can! WebGL, WebGPU, WebNN, Experimental Prompt() API