

# Hybrid AI for the Web

Michael McCool, Geoff Gustafson,  
Sudeep Divakaran, Muthaiah Venkatachalam

Intel

# Overview

- Current Status
- Specific Issues
- Goals and Requirements
- Questions for Discussion
- Proposed Next Steps

# Current Status

- WebNN use cases highlight advantages of client AI execution
- However, there are problems
  - Large models need long download times
  - Downloading a model and only then finding out it won't run
  - Startup time can be significant even after download
  - Sharing resources between multiple apps is difficult
  - Handling variation in capability among clients is difficult
  - No good way to optimize a model for specific client capabilities

# Specific Issues

- Model Management
  - Large open models cannot be reused across origins
  - Model storage and management opaque to the user
  - Cache eviction may not match user preferences
- Elasticity through Hybrid AI
  - Distributing work between client and server
  - Difficult to predict performance on a specific client
  - Sharing client capability details is a privacy risk
- User Experience
  - Privacy behavior may be unclear and may not match user preferences
  - Managing latency of model downloads

# Goals and Requirements

- Maximize ease of use *for the end user*
  - Minimize load times and meet latency targets
- Portability and elasticity
  - Minimize costs
  - Support clients of varying capabilities
  - Adapt based on resource availability
- Data privacy
  - Personal and business data
  - Support user choice and control
- Developer ease of use and consistency

# Questions for Discussion

How to...

- Handle model download latency and storage?
- Match model requirements to client capabilities?
- Choose among model fidelity levels?
- Support progressive transmission of models?
- Partition single models, support separate models, or both?

What are the priorities?

Do we need specific use cases for hybrid AI?

# Proposed Next Steps

1. Make sure we are solving the right problem
  - Feedback on proposed explainer
  - <https://github.com/webmachinelearning/proposals/issues/5>
2. Build a prototype implementation
  - e.g. Using the Model Loader API as a basis
  - We do have concrete ideas on how to solve the issues noted...
3. Bring back to group to discuss further