MELTing point

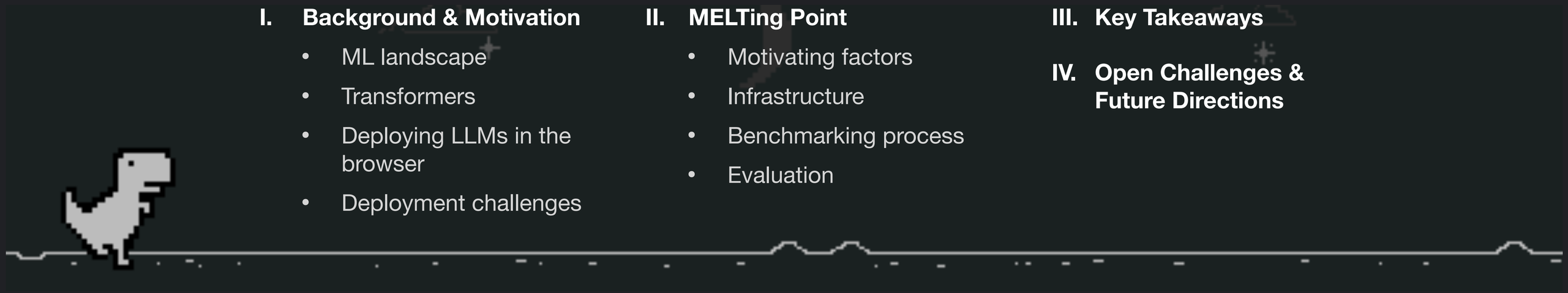
Mobile Evaluation of Language Transformers

Stefanos Laskaridis
ML Researcher @ Brave
mail@stefanos.cc
https://stefanos.cc

22nd July 2024

PEARG - IETF 120

Index



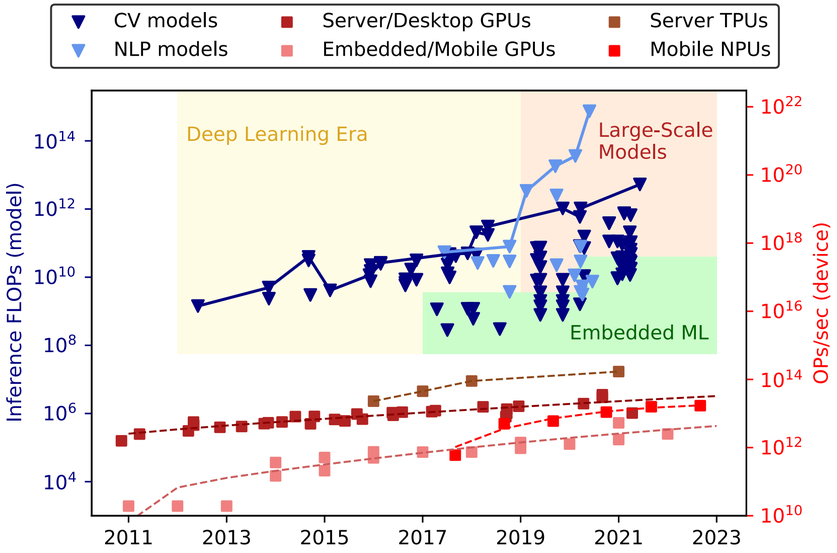
I. **Background & Motivation**
   - ML landscape
   - Transformers
   - Deploying LLMs in the browser
   - Deployment challenges

II. **MELTing Point**
   - Motivating factors
   - Infrastructure
   - Benchmarking process
   - Evaluation

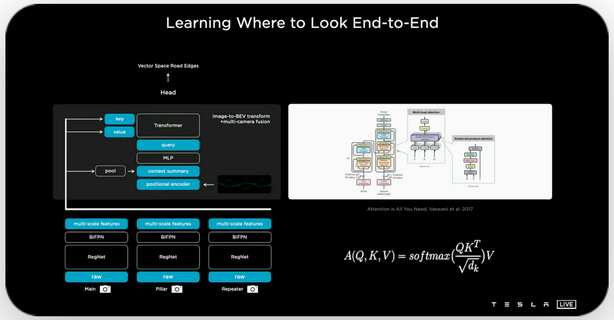
III. **Key Takeaways**

IV. **Open Challenges & Future Directions**

ML Landscape



DNNs are getting **more parameters** and become **evermore accurate** [1]



Self-driving
(Tesla FSD)

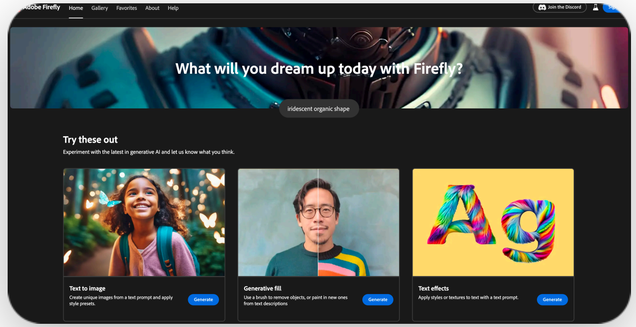
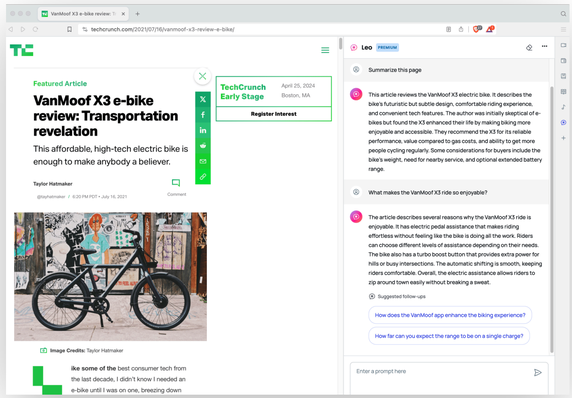
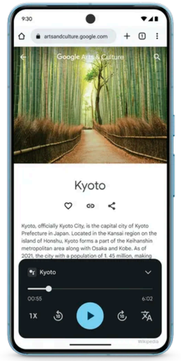


Image in/out-painting
(Adobe Firefly)



Webpage Interaction
(Brave Leo)

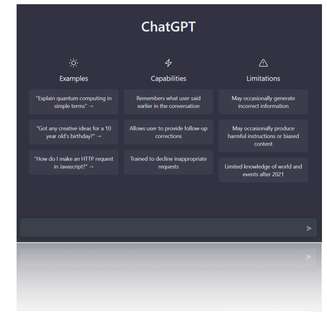


Voice Assistant
(Google Assistant on Pixel 8)

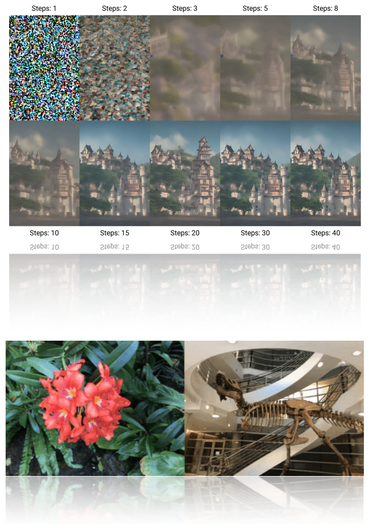
1. **Laskaridis, S.**, Venieris, S. I., Kouris, A., Li, R., & Lane, N. D. (2024). The Future of Consumer Edge-AI Computing. IEEE Pervasive Computing

ML Deployment Trends

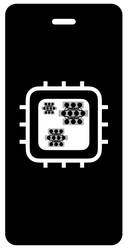
New architectures in the block:
LLMs [1], Diffusion Models [2], NERFs [3], SSMs [4]



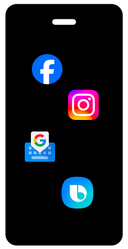




Multiple DNNs
residing on the device [5]

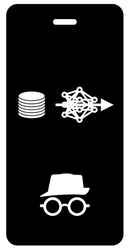


In the same app …

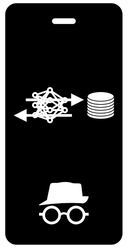


… or across apps

Need for Privacy-preserving Machine Learning (PPML)



Local private inference [6]



On-device training and personalisation [7]

1. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
2. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).
3. Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1), 99-106.
4. Gu, Albert, and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces." arXiv preprint arXiv:2312.00752 (2023).
5. Venieris, Stylianos I., Christos-Savvas Bouganis, and Nicholas D. Lane. "Multi-DNN Accelerators for Next-Generation AI Systems." arXiv preprint arXiv:2205.09376 (2022).
6. Ponomareva, Natalia, et al. "How to dp-fy ml: A practical guide to machine learning with differential privacy." Journal of Artificial Intelligence Research 77 (2023): 1113-1201.
7. Bell, J. H., Bonawitz, K. A., Gascón, A., Lepoint, T., & Raykova, M. (2020, October). Secure single-server aggregation with (poly) logarithmic overhead. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications

Transformers: Attention is all you need

- DNNs come in various forms and footprints
- Transformers are a novel architecture, introduced by Google, able to:
  - Attend to **long-context dependencies**
  - **Parallelize** workload tractably well
  - Model **various modalities**
    (text, speech, vision)
  - **Scale to large sizes** without performance plateauing.



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

arXiv:1706.03762v5 [cs.CL] 6 Dec 2017

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformers: Attention is all you need

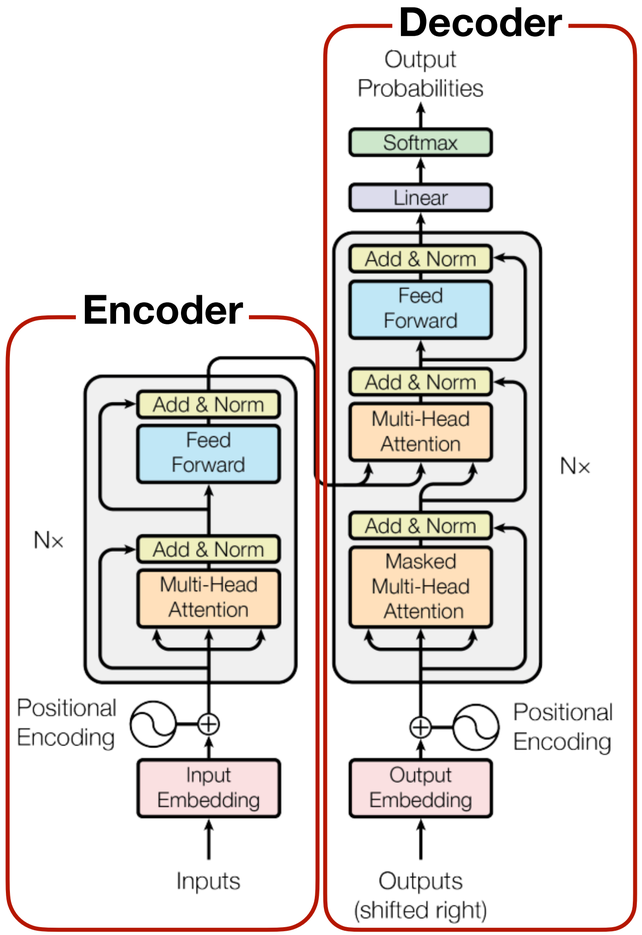
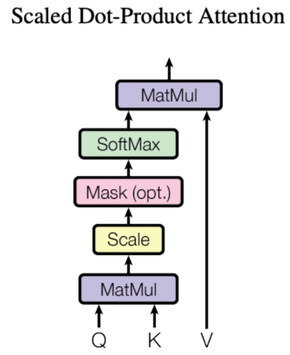
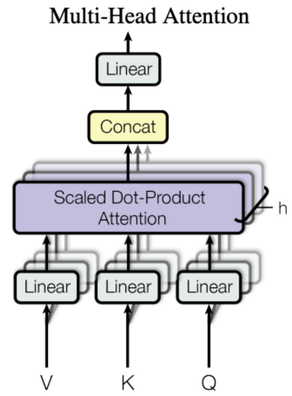


Figure 1: The Transformer - model architecture.

- Task: Given a sequence of tokens, predict the next token.
- Transformer enables the extraction and attendance to information from the whole sequence.
- Important entities:
  - Keys (K)
  - Values (V)
  - Queries (Q)





$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformers: vol. 2

- Decoder is *auto-regressive*.
  - $X = \{x_1, x_2, \ldots, x_t\}$, model generates x_{t+1}
- *Key-Value cache* optimises by storing intermediate attention states.
- Operations:
  - *"Prefill"* fills the cache, compute-bound operation
  - *"Generate"* produces token x_{t+1}, memory-bound operation

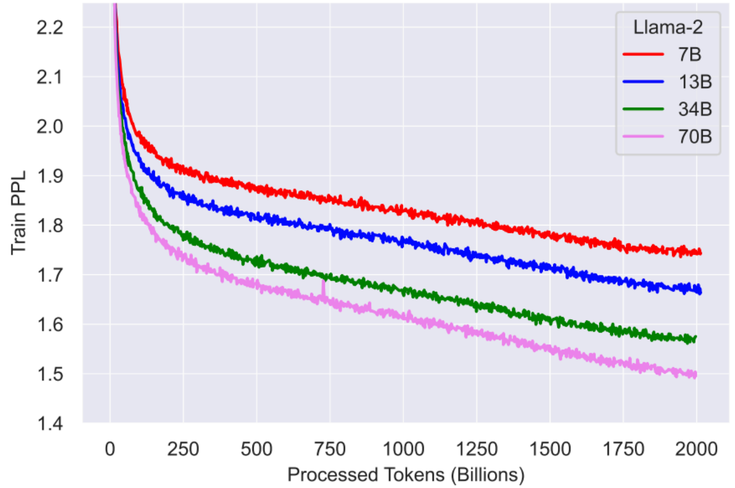
Terminology:



- **Context size:** Maximal window of tokens transformer attends to.
- **Max generated length:** Maximum generated tokens in output.
- **<EOS>:** end-of-sequence token

Large Language Models

- Transformers scale very well without saturating

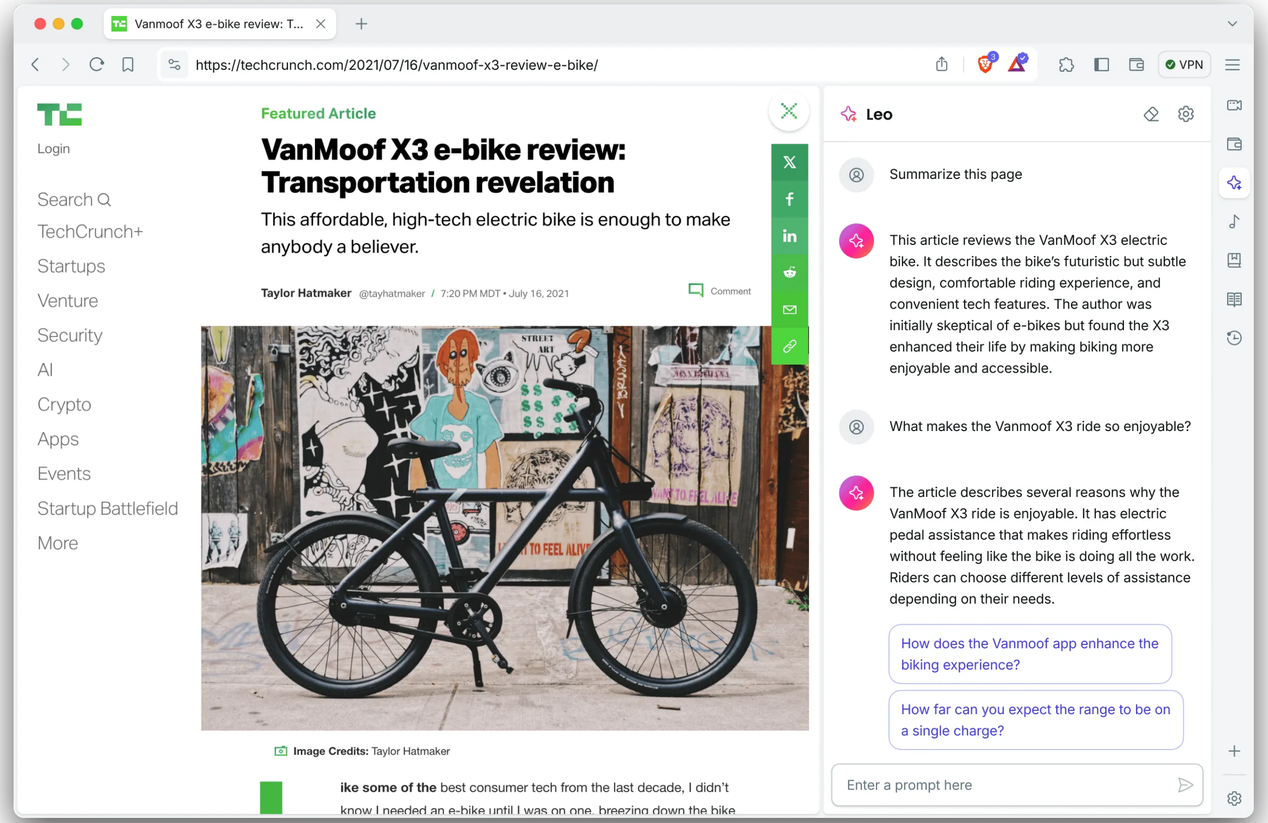


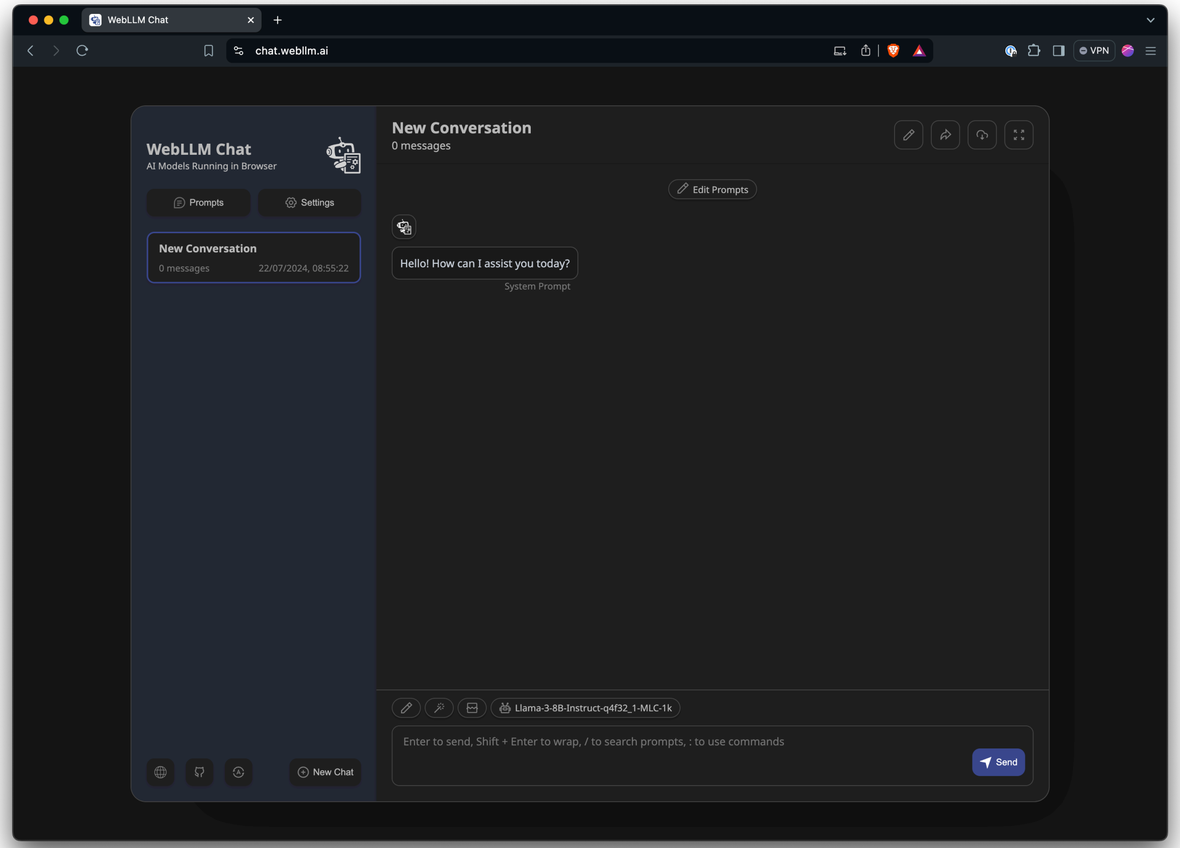
- Various players in the game, including startups and long-established organizations.
- **Accuracy**, **efficiency** and **cost** are key characteristics
- Multi-modality (e.g. ability to handle images)
- Integrations (e.g. agents/plug-ins, training on premise)

- Training process:
  - Pre-training on large corpora
  - Fine-tuning (PEFT, (Q)LoRA, etc.)
  - RLHF for alignment to human expectations

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Deploying LLMs in the Browser





Application-level

Website-level

LLMs in the Browser: Challenges

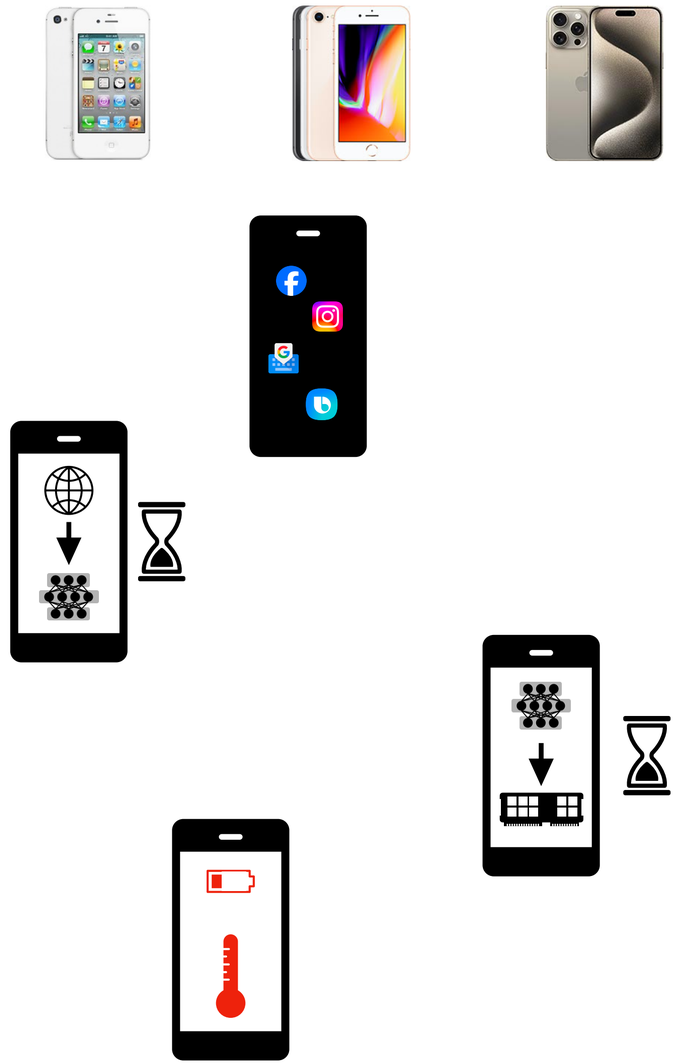
- **System heterogeneity**
  Low, mid and high tier devices co-exist out-in-the wild, along with previous generation devices..

- **Simultaneous workloads**
  An LLM is probably not the only task that is running on device. Simultaneous workloads may be neural or not.

- **Download latency**
  Model parameters take a noticeable time to distribute.

- **Initialization latency**
  Loading model weights from memory takes a noticeable time. Keeping in memory consumes power and non-negligible resources.

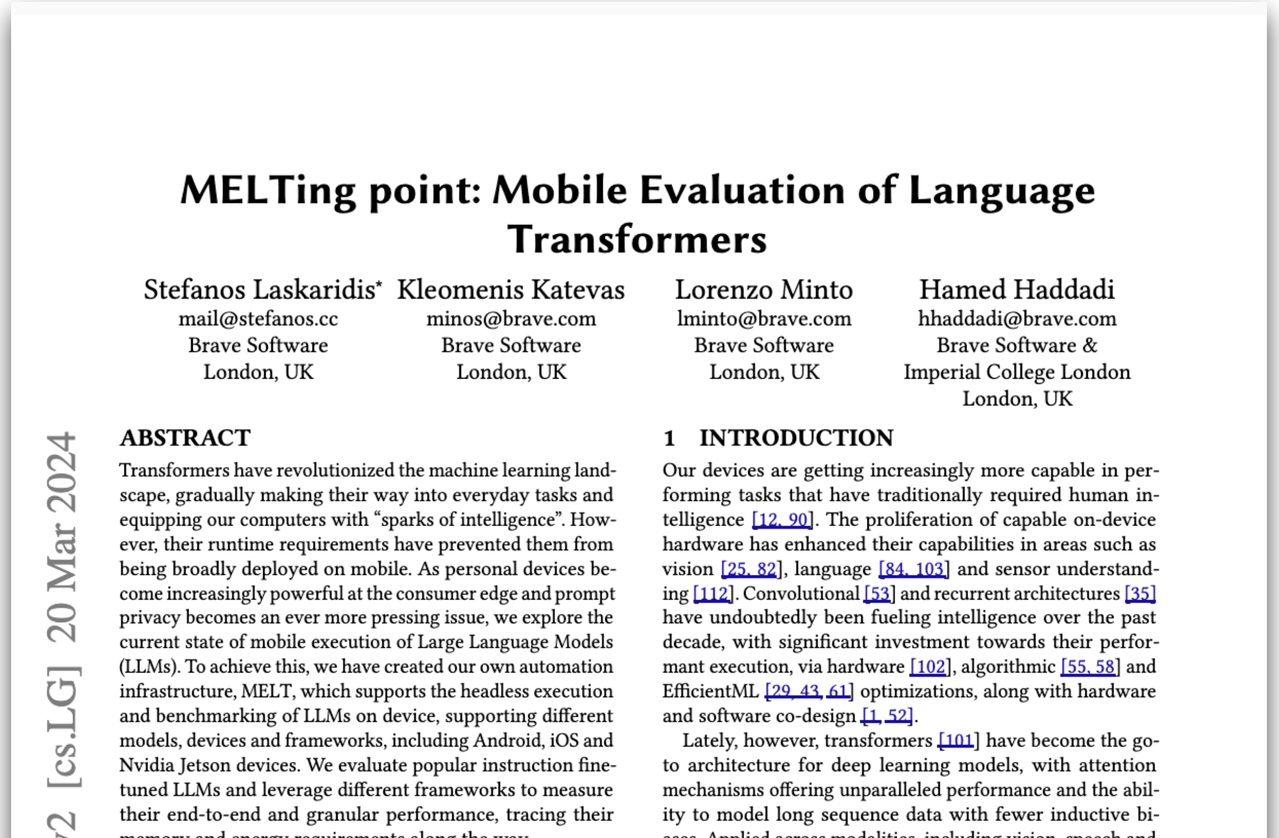
- **Sustained performance**
  Peak performance may not be guaranteed under continuous inference.



Accepted @ MobiCom'24

MELTing point

Mobile Evaluation of Language Transformers



MELTing point: Mobile Evaluation of Language Transformers

Stefanos Laskaridis* mail@stefanos.cc Brave Software London, UK
Kleomenis Katevas minos@brave.com Brave Software London, UK
Lorenzo Minto lminto@brave.com Brave Software London, UK
Hamed Haddadi hhaddadi@brave.com Brave Software & Imperial College London London, UK

2 [cs.LG] 20 Mar 2024

ABSTRACT

Transformers have revolutionized the machine learning landscape, gradually making their way into everyday tasks and equipping our computers with "sparks of intelligence". However, their runtime requirements have prevented them from being broadly deployed on mobile. As personal devices become increasingly powerful at the consumer edge and prompt privacy becomes an ever more pressing issue, we explore the current state of mobile execution of Large Language Models (LLMs). To achieve this, we have created our own automation infrastructure, MELT, which supports the headless execution and benchmarking of LLMs on device, supporting different models, devices and frameworks, including Android, iOS and Nvidia Jetson devices. We evaluate popular instruction fine-tuned LLMs and leverage different frameworks to measure their end-to-end and granular performance, tracing their

1 INTRODUCTION

Our devices are getting increasingly more capable in performing tasks that have traditionally required human intelligence [12, 90]. The proliferation of capable on-device hardware has enhanced their capabilities in areas such as vision [25, 82], language [84, 103] and sensor understanding [112]. Convolutional [53] and recurrent architectures [35] have undoubtedly been fueling intelligence over the past decade, with significant investment towards their performant execution, via hardware [102], algorithmic [55, 58] and EfficientML [29, 43, 61] optimizations, along with hardware and software co-design [1, 52].

Lately, however, transformers [101] have become the go-to architecture for deep learning models, with attention mechanisms offering unparalleled performance and the ability to model long sequence data with fewer inductive bi-

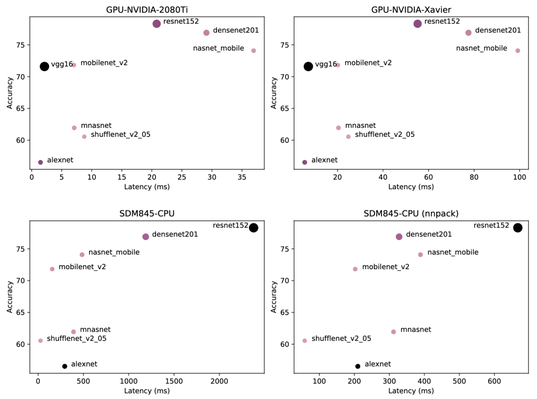
arXiv.2403.12844

https://github.com/brave-experiments/MELT-public

Previous Work

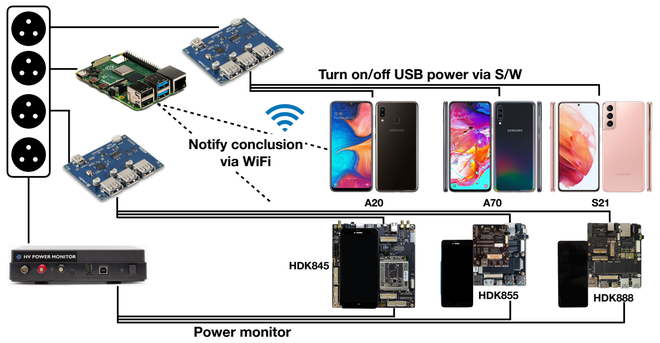
EmBench [EMDL'20]

Controlled study of the performance of SOTA DNNs (circa 2018) across different devices.



gaugeNN [IMC'21]

Study of the performance of in-the-wild DNNs (found in top apps) across different devices.



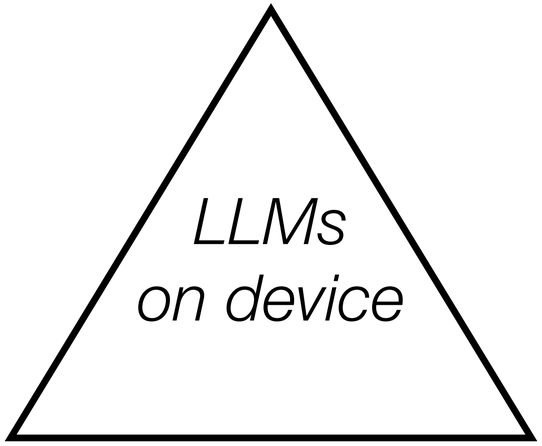
Almeida, M., **Laskaridis, S.**, Leontiadis, I., Venieris, S. I., & Lane, N. D. 2019. EmBench: Quantifying performance variations of deep neural networks across modern commodity devices. In The 3rd international workshop on deep learning for mobile systems and applications (pp. 1-6).

Almeida, M., **Laskaridis, S.**, Mehrotra, A., Dudziak, L., Leontiadis, I., & Lane, N. D. (2021, November). Smart at what cost? characterising mobile deep neural networks in the wild. In Proceedings of the 21st ACM Internet Measurement Conference (pp. 658-672).

Why deploy on device?

De-centralization & Democratisation





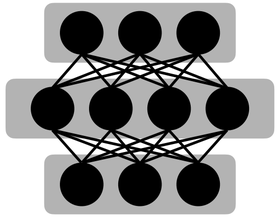


Privacy & Personalisation

Sustainability

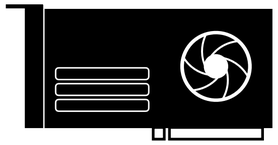


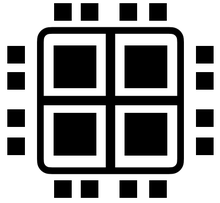
From LLMs to SLMs



LLMs

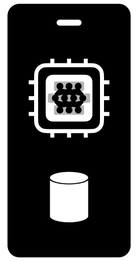
Models: Llama-3, Mixtral, DBRX, Command-R+
Frameworks: GPT-NeoX, DeepSpeed, TGI
Deployment: Datacenter hosting





SLMs

Models: Phi-3, Gemma2, TinyLlama, MobileLLM
Frameworks: llama.cpp, MLC, MediaPipe, MLX
Deployment: Edge or on-device



https://github.com/stevelaskaridis/awesome-mobile-llm

The MELT Infrastructure

