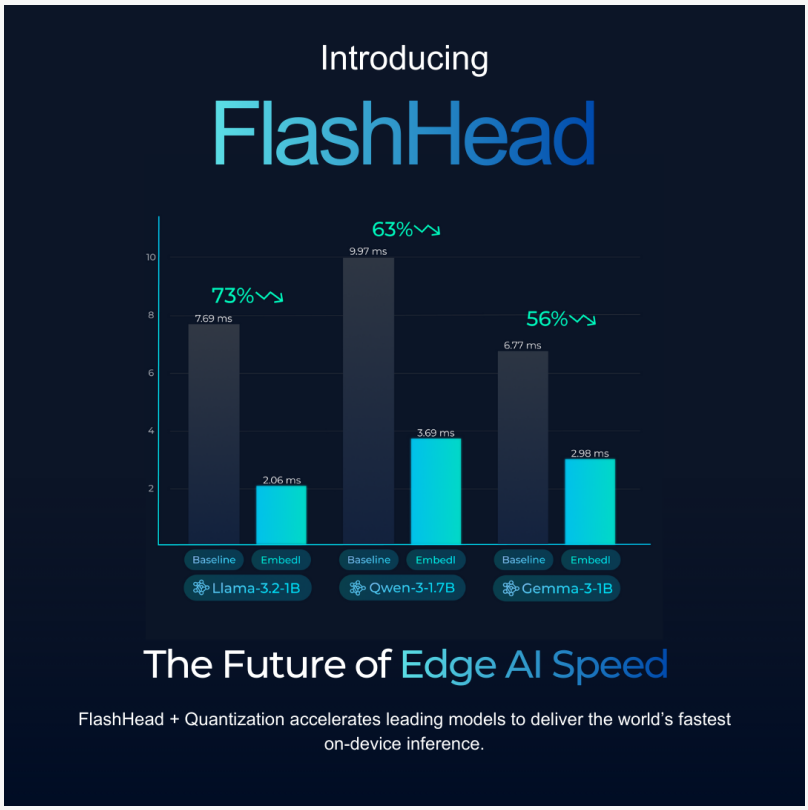
Embedl sets a new standard for on-device LLM inference, releasing the world's fastest language models for the edge

We announce FlashHead, a technical breakthrough that makes Llama-3.2, Gemma-3, and Qwen-3 the world's fastest models for on-device inference.



FlashHead + Quantization accelerates leading models to deliver the world's fastest on-device inference.

Embedl

GOTHENBURG, SWEDEN, December 8, 2025 /EINPresswire.com/ -- Embedl, a Swedish deep-tech pioneer in AI model optimization, announced today FlashHead, an optimization method that makes the most popular language models, including Llama-3.2 (Meta), Gemma-3 (Google DeepMind), and Qwen-3 (Alibaba), the fastest models for on-device inference.

The technology, "FlashHead: Efficient Drop-in Replacement for the Classification Head in Language Model Inference," reduces latency by up to 43% while preserving full model accuracy.

"FlashHead eliminates a major AI deployment bottleneck," says Hans Salomonsson, CEO of Embedl. " This means world class SLM models now run at lightning speed on everyday devices; fast, compact, and sustainable."

A Revolution in Language Model Efficiency

Today's classification head predicts the next token by assigning a probability to all possible tokens, but it is resource-intensive. State-of-the-art models include hundreds of thousands of

possible tokens in their vocabularies, causing the head to become a severe bottleneck for inference. FlashHead reformulates a language model through the lens of information retrieval, making it faster and less computationally demanding. Notably, FlashHead delivers these efficiency gains while leaving the model's output virtually unchanged.

FlashHead achieves this through several innovations, including equal-sized clustering for fast memory access, multi-probe retrieval, probabilistic sampling for increased speed, and selective quantization.

1. Equal-sized clustering for fast memory access.
2. Multi-probe retrieval to evaluate multiple token clusters efficiently.
3. Probabilistic probe sampling for accurate, high-speed decoding.
4. Selective quantization for robust low-bit computation without accuracy loss.

These optimizations make the classification head a minor cost, enabling fast inference even on low-power devices.

Proven Across Model Families
FlashHead has been tested on several of the world's most widely used open models:

- Llama-3.2-1B (Meta) 43% Latency Reduction
- Gemma-3-270M (Google DeepMind) 26% Latency Reduction
- Qwen-3-1.7B (Alibaba) 24% Latency Reduction

These speedups are in relation to state-of-the-art optimization (w4A16 quantization). When combined with mixed precision optimization, the Embedl optimized Llama 3.2-1B for RTX 3500 Ada Generation reaches almost the same latency (2.06 ms) as the original model on the much more powerful H200 GPU (1.95 ms),, enabling true on-device AI.

Available December 8 on Hugging Face
Starting December 8, 2025, developers and researchers can access and use optimized FlashHead models for Llama-3.2, Gemma-3, and Qwen-3 on Hugging Face. Visit Hugging Face to try FlashHead and experience its efficiency improvements firsthand.

"This milestone makes state-of-the-art models run faster, cheaper, and locally for everyone, no cloud required," says Hans Salomonsson.

Frida Dygd Horwath
Embedl
frida@embedl.com

This press release can be viewed online at: https://www.einpresswire.com/article/873620991

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors

try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.
© 1995-2025 Newsmatics Inc. All Right Reserved.