Contribution ID: **19** Type: **not specified**

Statistical Evaluation of CPU-Based Offline LLMs for Industrial Text Processing on Low-Cost Edge Hardware

Recent advances in generative AI have enabled powerful language models for industrial applications. However, most solutions rely on cloud-based infrastructures or GPU-accelerated environments, which raise concerns regarding data privacy, latency, and operational cost—particularly in industrial settings dealing with sensitive internal documents.

In this study, we investigate the feasibility of deploying offline large language models (LLMs) on CPU-only edge hardware, such as standard notebooks and low-cost mini PCs. In particular, we evaluate the performance of highly quantized models, including emerging architectures such as BitNet, within a real-world industrial use case.

The application scenario is based on a production system developed using n8n, where incoming customer emails are processed locally without external data transfer. Two core tasks are considered:

1. Structured information extraction (classification task): Extraction of machine identifiers and request types from customer emails.
2. Text summarization and interpretation (generation task): Generation of concise summaries and actionable insights from unstructured text.

For the classification task, performance is evaluated using classical statistical metrics derived from the confusion matrix, including precision, recall, and F1-score. For the generative task, we apply G-Eval-based metrics to assess dimensions such as correctness, completeness, and consistency.

Beyond output quality, we introduce a comprehensive set of system-level performance indicators, including:

- processing latency
- tokens generated per second
- total token count per task
- CPU utilization and memory footprint

These metrics are analyzed under different deployment configurations, comparing execution on a high-performance notebook (64 GB RAM) and a low-cost edge device (Intel N95 mini PC, 12 GB RAM).

Our results demonstrate that CPU-based offline LLMs can achieve competitive performance for industrial text processing tasks, while significantly reducing infrastructure complexity and enabling fully local data processing. The study highlights that, with appropriate model selection and evaluation metrics, cost-efficient edge deployments without GPUs are a viable alternative for industrial AI applications.

Special/ Invited session

Classification

Mainly application

Keywords

Edge Computing, Quantized Language Models (BitNet), Offline-LLMs

Primary author: Dr MOESER, Guido (masem research institute)

Co-authors: AHLEMEYER-STUBBE, Andrea (Data Mining & More); DUB, Igor (Wiesbaden Smart City Department); BUMM, Rainer (PHOENIX)

Presenter: Dr MOESER, Guido (masem research institute)

Track Classification: Trustworthy and Explainable AI