Reduce your cloud OPEX for AI/ML services with Latent AI

Organizations need a solution to lower their cloud operational expenses (OPEX) while producing better models more quickly. The Latent AI Efficient Inference Platform (LEIP) is an MLOps SDK, an innovative software technology that significantly lowers AI/ML inference cost by optimizing the model runtime.

LEIP on the cloud

**40% IN COST SAVINGS**

AWS & Latent AI jointly sponsored a third party benchmark study comparing Amazon SageMaker workflow with and without LEIP optimization.

```
<table>
<thead>
<tr>
<th>Inference Cost ($/million in.)</th>
<th>SageMaker</th>
<th>LEIP Optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$8.30</td>
<td>$5.20</td>
</tr>
</tbody>
</table>
```

*Saved 40% on OPEX (improved inference with less storage)*

Model Architecture - YoloV5 Large
Inference Target - G4Dn.xlarge

LEIP offers a dedicated and automated means to lower cloud costs.

With LEIP, customers can:

- Reduce OPEX without changing current SageMaker workflow
- Increase inference speed while lowering inference costs
- Develop and deploy models repeatedly with guaranteed runtime
- Improve time to market with a dedicated edge MLOps pipeline

**48% FASTER INFEERENCE**

```
<table>
<thead>
<tr>
<th>Model Accuracy (mAP)</th>
<th>SageMaker</th>
<th>LEIP Optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>44.61</td>
<td>40.61</td>
</tr>
</tbody>
</table>
```

Model Architecture - YoloV5 Large
Inference Target - G4Dn.xlarge

Ready to integrate LEIP with your SageMaker workflow? Visit latentai.com or contact mllops@latentai.com today!