

Intel® GenAI Hackathon

Team Name : TEAM VELOCE

Problem Statement : AI FOR SOCIAL IMPACT

AI FOR SOCIAL IMPACT

Challenge participants to use Stable Diffusion and Transformers to develop AI models that can address social issues. This could involve predicting areas at risk of natural disasters, generating educational content for underprivileged students, etc. The focus here is on open innovation for social good.



UNIQUE IDEA

The Indo-LLaMA tool interprets the text while keeping the essence of what is to be conveyed.

To achieve this the tool scans through {paragraphs → lines} rather than {word → word} interpretation.

This keeps the meaning intact. The tool also provides a summary of the content provides to get a jist of what is to be conveyed.

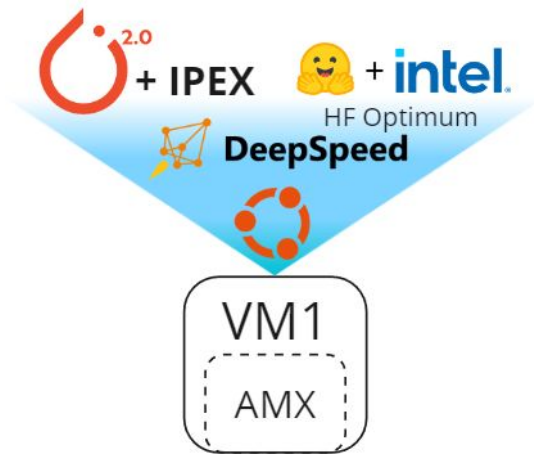
ARTIFICIAL RESPONSIBLE
INTELLIGENCE?

LETS TALK
LLAMA

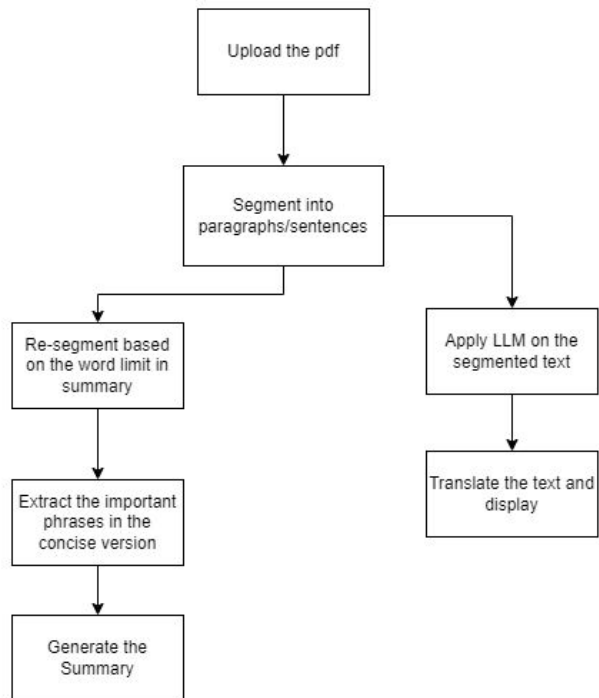


FEATURES OFFERED

- Translating from and to multiple national/ international languages.
- Inclusion of new languages due to presence of independent tokenizer code
- Provision of a concise summary to understand the gist of the content
- Due to uses of INTEL's Xeon VM, computation time is reduced.
- Large volume of data can be stored on INTEL cloud.

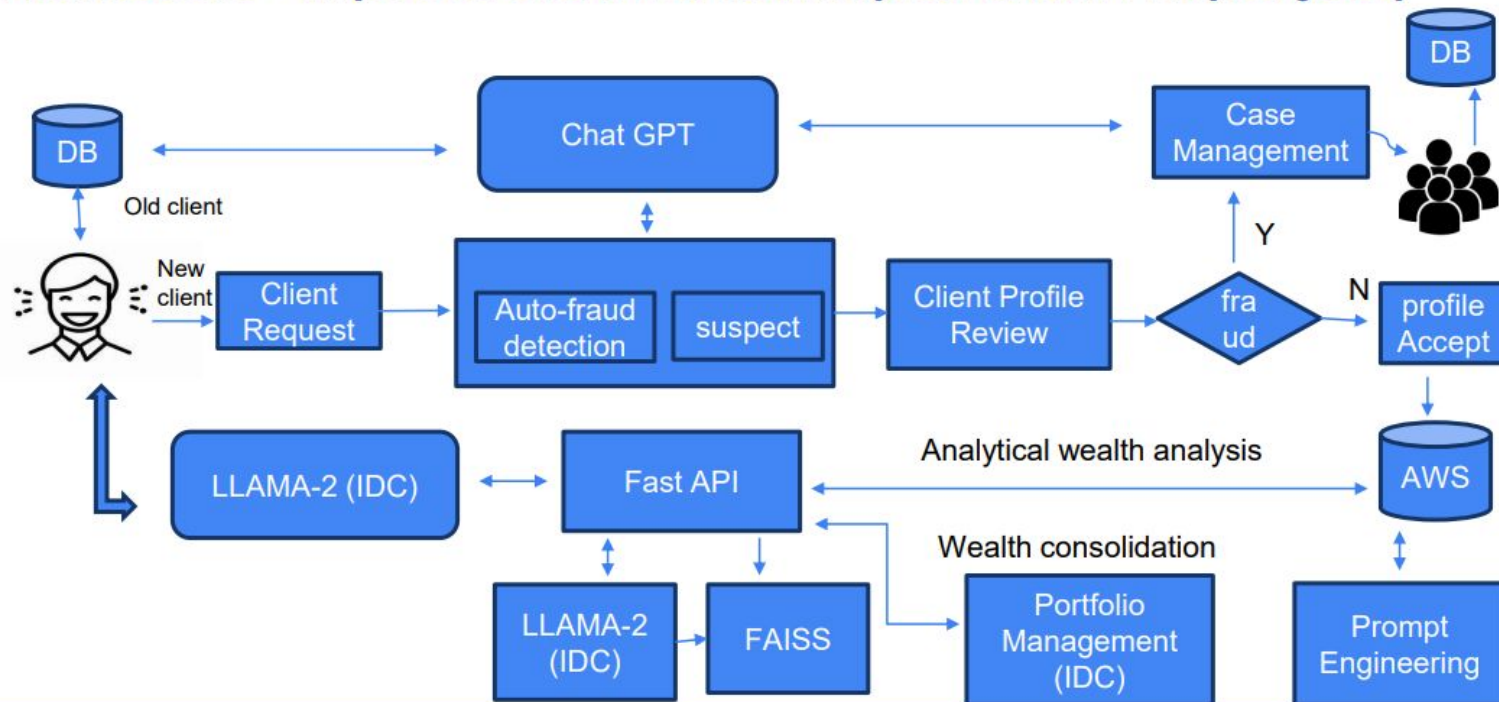


PROCESS FLOW



ARCHITECTURE DIAGRAM

Architecture – Impact of oneAPI AI Toolkit (How oneAPI helped you?)



TECHNOLOGIES USED

- *intel-extension-for-pytorch : ipex*

- *Tensorflow*

- *Transformers*

- *deep_translator*

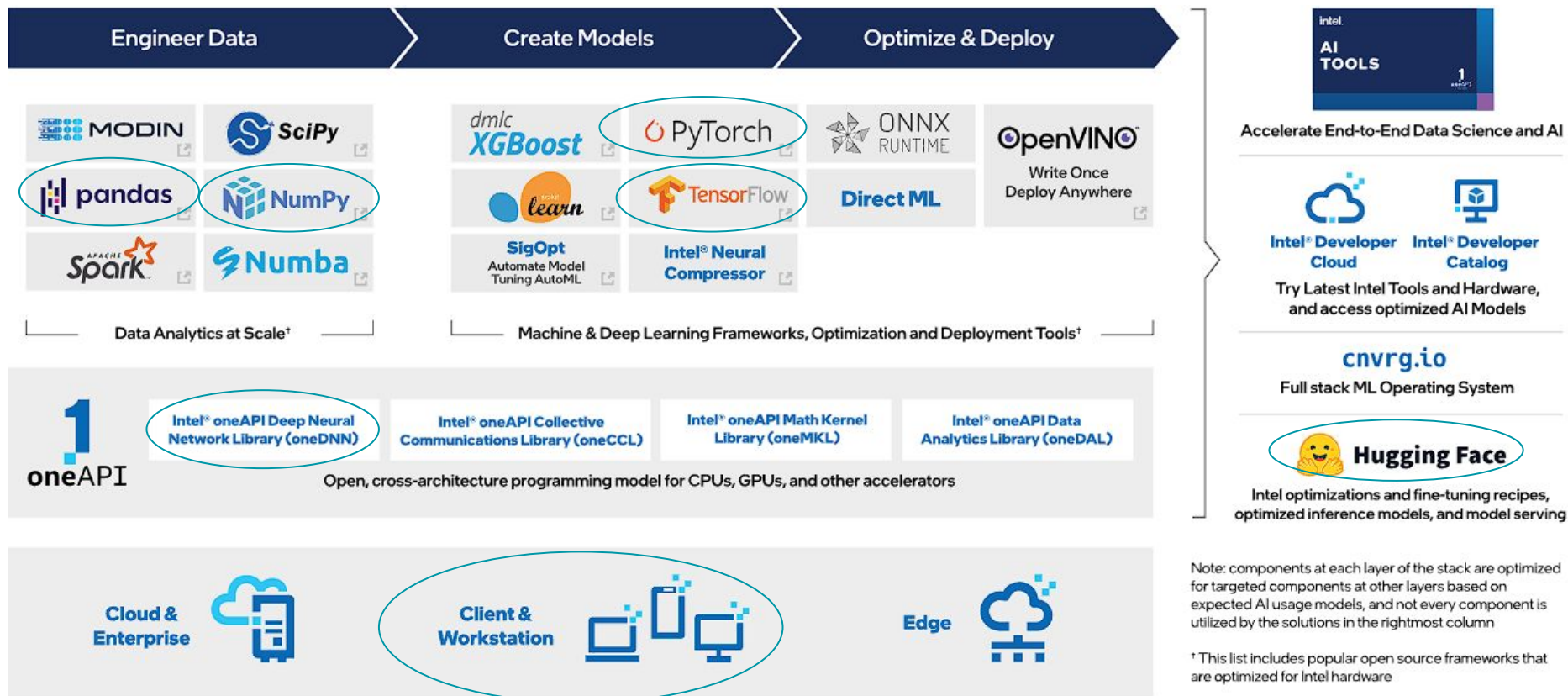
Both pytorch and ipex are used due to dependency issues.



Transformers is a library exclusively for interacting with models on hugging face

At this stage we use these libraries primarily but on development of our own LLM's with large amount of data, the other tools will come to be useful

SCALABILITY



Intel® Developer Cloud Account (Screenshot)

Account Settings

Your Developer Cloud Account

Cloud Account ID: 826312421446

Tier: Standard

Upgrade

Standard tier includes:

- Explore and evaluate the latest Intel® AI products.
- Develop AI skills.
- Access cutting edge learning resources.
- Get support from the Intel community.

Not included:

- Early prerelease hardware
- AI and machine learning software toolkits
- Billed subscription for teams
- Use CPU, GPU, and AI accelerators
- Intel premium support

Your intel.com Account

Name:

Devesh Shreyas

Email:

shreyasdev154@gmail.com

Use case of Intel® Developer Cloud (IDC)

- **Evaluating and optimizing performance:**
 - IDC provides access to a variety of powerful Intel Xeon processors and GPUs, allowing you to benchmark INDO LLAMA on different hardware configurations and optimize its performance for specific use cases. This can lead to faster processing times and improved efficiency.
- **Scaling development and testing:**
 - IDC's cloud-based nature allows you to easily scale your development and testing environment up or down as needed. This is crucial for handling large datasets or complex text analysis tasks without worrying about hardware limitations.
- **Experimenting with advanced AI frameworks:**
 - IDC offers pre-installed deep learning frameworks like TensorFlow and PyTorch, which INDO LLAMA can leverage for advanced text analysis tasks like sentiment analysis or topic modeling. You can experiment with different frameworks and configurations to find the best fit for your specific needs.

EFFICIENCY AND EFFECTIVENESS BENCHMARK


```
[4] model = MBartForConditionalGeneration.from_pretrained("facebook/mbart-large-50-one-to-many-mmt")
```


pytorch_model.bin: 100% ██████████ 2.44G/2.44G [00:38<00:00, 32.6MB/s]


/usr/local/lib/python3.10/dist-packages/torch/_utils.py:831: UserWarning: TypedStorage is deprecated
return self.fget.__get__(instance, owner)()

generation_config.json: 100% ██████████ 261/261 [00:00<00:00, 14.8kB/s]

```
[5] tokenizer = MBart50TokenizerFast.from_pretrained("facebook/mbart-large-50-one-to-many-mmt", src_
```

tokenizer_config.json: 100%  528/528 [00:00<00:00, 28.0kB/s]

sentencepiece.bpe.model: 100%  5.07M/5.07M [00:00<00:00, 19.8MB/s]

special_tokens_map.json: 100%  717/717 [00:00<00:00, 42.0kB/s]

```
[10]
# translate from English to Hindi
generated_tokens = model.generate(
    **model_inputs,
    forced_bos_token_id=tokenizer.lang_code_to_id["hi_IN"]
)
```

$$T_{\text{total}} = 205\text{s}$$

Run on Google Colab Runtime

```
In [14]: %%timeit
from transformers import MBartForConditionalGeneration, MBart50TokenizerFast
import torch
from intel_extension_for_pytorch.transformers.optimize import optimize

1.01 µs ± 118 ns per loop (mean ± std. dev. of 7 runs, 1,000,000 loops each)
```

```
In [15]: %%timeit
model = MBartForConditionalGeneration.from_pretrained("facebook/mbart-large-50-one-to-many-mmt")
tokenizer = MBart50TokenizerFast.from_pretrained("facebook/mbart-large-50-one-to-many-mmt", src_lang="en_XX")

2.18 s ± 638 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

```
In [16]: %%timeit
article_en = "Morality or moral behaviour is not necessarily the result of philosophical reflections. These philosophical reflections are not necessarily the result of philosophical reflections."
model_inputs = tokenizer(article_en, return_tensors="pt")

266 µs ± 1.93 µs per loop (mean ± std. dev. of 7 runs, 1,000 loops each)
```

```
In [17]: %%timeit
# translate from English to Hindi
generated_tokens = model.generate(
    **model_inputs,
    forced_bos_token_id=tokenizer.lang_code_to_id["hi_IN"]
)

12.9 s ± 737 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

```
In [18]: %%timeit
translation = tokenizer.batch_decode(generated_tokens, skip_special_tokens=True)

71.3 µs ± 13.2 µs per loop (mean ± std. dev. of 7 runs, 10,000 loops each)
```

```
In [20]: %%timeit
translation

8.18 ns ± 2.14 ns per loop (mean ± std. dev. of 7 runs, 100,000,000 loops each)
```

Run on Intel Xeon VM and
optimised libraries.

$$T_{\text{total}} = 40\text{s}$$

$$\eta = 80.4\% \uparrow$$

Languages offered

National

Gujarati (gu_IN), Hindi (hi_IN) Bengali (bn_IN), Malayalam (ml_IN), Marathi (mr_IN), Tamil (ta_IN), Telugu (te_IN),

International

Arabic (ar_AR), Czech (cs_CZ), German (de_DE), English (en_XX), Spanish (es_XX), Estonian (et_EE), Finnish (fi_FI), French (fr_XX), Italian (it_IT), Japanese (ja_XX), Kazakh (kk_KZ), Korean (ko_KR), Lithuanian (lt_LT), Latvian (lv_LV), Burmese (my_MM), Nepali (ne_NP), Dutch (nl_XX), Romanian (ro_RO), Russian (ru_RU), Sinhala (si_LK), Turkish (tr_TR), Vietnamese (vi_VN), Chinese (zh_CN), Afrikaans (af_ZA), Azerbaijani (az_AZ), Persian (fa_IR), Hebrew (he_IL), Croatian (hr_HR), Indonesian (id_ID), Georgian (ka_GE), Khmer (km_KH), Macedonian (mk_MK), Mongolian (mn_MN), Polish (pl_PL), Pashto (ps_AF), Portuguese (pt_XX), Swedish (sv_SE), Swahili (sw_KE), Thai (th_TH), Tagalog (tl_XX), Ukrainian (uk_UA), Urdu (ur_PK), Xhosa (xh_ZA), Galician (gl_ES), Slovene (sl_SI)

translated

परचक्रवर्ती राजा है, गुजराती नाम है। १९७८ में जो इनको ब्रिटिश के नाम से बमरुदी नाम दिया, वह १९७० के अंत और १९८० के आरंभ की क्रांति के दौरान बनी थी। क्रांति के दौरान ही ब्रिटिश, वायव्यीय वायव्यीय और वायव्यीय अर्थात् दमगुर्त अर्थात् इनकी संस्कृति के सम्पूर्ण कला है। १९८५ में, भारतीयों गुजरात सेवा एजेंसी में दो बमों को लाया और जहाज को दुबा, एक सदस्य मार दिया गया, लेकिन क्रांति आता और कुछ ही महीने ब्रिटिश को रोक नहीं सका था। गुजराती नाम में उसे एक नाम दिया है दमगुर्तवासी ब्रिटिश, और २२ वें को दमगुर्त, दमगुर्त, [अब मैंने ऐसा ही कहें: जो ब्रिटिशों द्वारा आयात प्रवासन और जहाज के नामों के नेकाओं पर होता है।] [नम नम राजे, राजे, राजे] [३]

डा. एम. राशमी व्याख्यान देते हैं कि कैसे जरूरतमंदों की सहायता की जा सकती है। दान एक सर्वेच्छक कार्य का एक रूप है जिस सीधे या अपरत्यक्ष रूप से लोगों की व्यक्तिगत या सामूहिक रूप से सहायता करने या सेवा प्रदान करने के लिए किया जा सकता है। गरीनपीस एक अंतरराष्ट्रीय गैर-सरकारी संगठन या गैर-सरकारी संगठन है जिसके कार्यालय 55 से अधिक देशों में हैं। गैर-सरकारी संगठन सामान्यतः ऐसे दान हैं जो एक देश में स्थित हैं लेकिन दूसरे देशों में काम करते हैं। वे अंतरराष्ट्रीय विकास के गैर-सरकारी संगठनों, आपदा राहत और मानवीय गैर-सरकारी संगठनों, शांति और मानवाधिकार के गैर-सरकारी संगठनों, संरक्षण के गैर-सरकारी संगठनों, बाल संरक्षण संगठनों जैसे श्रेणियों में आते हैं। गरीनपीस ने इसे एक नए जहाज से पुर्तस्थापित किया है दूसरा रैनबो वेयर 22 वर्ष

summary

Future Milestones

- Inclusion of more Indian languages.
 - This can be done by the 'translator.py'
- Upgrading to a more responsive website
- Training the model to understand various parts of speech in different languages.

Thank you so much!