User Query → Language Detection (8011) → Script Analysis

↓

If English → Pass through

If Indic + Roman → Transliterate (8010) → Translate (8011)

If Indic + Native → Translate (8011)

↓

Process with LLM (Fast/CrewAI)

↓

English Response → Translate back (8011) → Transliterate if needed (8010)

↓

Final Response in User's Language & Script

Models Used

For Language Identification: https://github.com/AI4Bharat/IndicLID

For Translation: https://github.com/ai4bharat/IndicTrans2

For transliteration Roman to Native: https://github.com/AI4Bharat/IndicXlit

For rule-based Transliteration: https://pypi.org/project/transliterate/

IndicLID

IndicLID, is a language identifier for all **22 Indian languages** listed in the Indian constitution in both **native-script and romanized text.** IndicLID is the first LID for romanized text in Indian languages. It is a two stage classifier that is ensemble of a fast linear classifier and a slower classifier finetuned from a pre-trained LM. It can predict 47 classes (24 native-script classes and 21 roman-script classes plus English and Others).

IndicTrans2

IndicTrans2 is the first open-source transformer-based multilingual NMT model that supports high-quality translations across all the **22 scheduled Indic languages** — including multiple scripts for low-resource languages like Kashmiri, Manipuri and Sindhi. It adopts script unification wherever feasible to leverage transfer learning by lexical sharing between languages. Overall, the model supports five scripts Perso-Arabic (Kashmiri, Sindhi, Urdu), Ol Chiki (Santali), Meitei (Manipuri), Latin (English), and Devanagari (used for all the remaining languages).

IndicXlit

IndicXlit is a transformer-based multilingual transliteration model (~11M) that supports **21 Indic languages** for Roman to native script and native to Roman script conversions. It is trained on Aksharantar dataset which is the largest publicly available parallel corpus containing 26 million word pairs spanning 20 Indic languages at the time of writing (5 May 2022). It supports the following 21 Indic languages:

Rule-based Transliteration

Languages Supported

No.	Language	IndicLID (Roman)	IndicLID (Native)	IndicXLit Code	IndicTrans2 Code(s)
1	Assamese	asm_Latn	asm_Beng	as	asm_Beng
2	Bengali	ben_Latn	ben_Beng	bn	ben_Beng
3	Bodo	brx_Latn	brx_Deva	brx	brx_Deva
4	Dogri	–	doi_Deva	–	doi_Deva
5	Konkani	kok_Latn	kok_Deva	gom	gom_Deva
6	Gujarati	guj_Latn	guj_Gujr	gu	guj_Gujr
7	Hindi	hin_Latn	hin_Deva	hi	hin_Deva
8	Kannada	kan_Latn	kan_Knda	kn	kan_Knda
9	Kashmiri	kas_Latn	kas_Arab / kas_Deva	ks	kas_Arab / kas_Deva
10	Maithili	mai_Latn	mai_Deva	mai	mai_Deva
11	Malayalam	mal_Latn	mal_Mlym	ml	mal_Mlym
12	Manipuri	mni_Latn	mni_Beng / mni_Meti	mni	mni_Beng / mni_Mtei
13	Marathi	mar_Latn	mar_Deva	mr	mar_Deva
14	Nepali	nep_Latn	nep_Deva	ne	npi_Deva
15	Odia	ori_Latn	ori_Orya	or	ory_Orya
16	Punjabi	pan_Latn	pan_Guru	pa	pan_Guru
17	Sanskrit	san_Latn	san_Deva	sa	san_Deva
18	Santali	–	sat_Olch	–	sat_Olck

19	Sindhi	snd_Latn	snd_Arab	sd	snd_Arab / snd_Deva
20	Sinhala	–	–	si	–
21	Tamil	tam_Latn	tam_Tamil	ta	tam_Taml
22	Telugu	tel_Latn	tel_Telu	te	tel_Telu
23	Urdu	urd_Latn	urd_Arab	ur	urd_Arab
24	English	eng_Latn	–	en	eng_Latn
	Other	–	other	–	–

Analysis

Dogri: https://en.wikipedia.org/wiki/Dogri_language

Dogri training required for IndicLID and IndicTrans2 for roman languages

Sinhala: https://en.wikipedia.org/wiki/Sinhala_language

Sinhala is removed for now.

Santali: https://en.wikipedia.org/wiki/Santali_language

Santali training required for IndicLID and IndicTrans2 for roman languages

Results Comparison

Language	Samples	IndicLID Accuracy	Roman->Native Similarity	Native->English Similarity	Top Misdetections
Gujarati	4	100.00%	97.89%	91.71%	[]

Marathi	4	100.00%	96.73%	80.56%	[]
Bengali	3	100.00%	96.52%	76.03%	[]
Bodo	3	0.00%	43.36%	40.54%	[('eng_Latn', 1), ('nep_Latn', 1)]
Dogri	3	0.00%	12.35%	45.46%	[('other', 1), ('kas_Latn', 1)]
Assamese	3	100.00%	94.18%	77.57%	[]
Hindi	3	66.67%	72.85%	65.07%	[('asm_Latn', 1)]
Kannada	3	100.00%	93.68%	80.55%	[]
Konkani	3	100.00%	96.34%	78.16%	[]
Kashmiri	3	100.00%	16.76%	51.15%	[]
Maithili	3	66.67%	84.30%	54.09%	[('brx_Latn', 1)]
Manipuri	3	100.00%	16.48%	64.48%	[]
Nepali	3	100.00%	95.35%	89.58%	[]

Odia	3	100.00%	93.40%	78.49%	[]
Sindhi	3	33.33%	48.30%	62.23%	[('nep_Latn', 1), ('brx_Latn', 1)]
Punjabi	3	66.67%	70.08%	62.44%	[('eng_Latn', 1)]
Sanskrit	3	100.00%	93.29%	67.33%	[]
Santali	3	0.00%	17.30%	21.31%	[('nep_Latn', 2), ('asm_Latn', 1)]
Telugu	3	100.00%	97.46%	74.38%	[]
Tamil	3	100.00%	95.92%	81.78%	[]
Urdu	3	33.33%	40.87%	36.12%	[('asm_Latn', 1), ('brx_Latn', 1)]
Malayalam	2	100.00%	96.17%	71.27%	[]

Evaluation Metrics

1. IndicLID Confidence

- What it is:
  This is the confidence score output by your language identification model (IndicLID).
  It predicts the language of the Roman input text and gives a probability (0 to 1).
  - Example: 0.92 → The model is 92% confident that the Roman text is in Hindi.
- Why it matters:
  If the wrong language is detected, all subsequent transliteration/translation steps will likely fail.

difflib.SequenceMatcher (Python Standard Library)

difflib.SequenceMatcher is a class in Python's built-in difflib module.
It is used to compare two sequences (strings, lists, etc.) and find how similar they are.

Key Points

- It finds the **longest contiguous matching subsequence** between two sequences.
- It computes a **similarity ratio** (between 0 and 1).
- Commonly used for:
  - String similarity
  - Diff utilities (showing differences between texts)
  - Approximate matching (e.g., "close" spellings)
  - It doesn't understand synonyms, grammar, or semantics.

  Example:

- "happy" vs "joyful" → difflib says similarity is **0** (completely different letters).
- But semantically, they mean the same.

Similarity Ratio Formula

ratio=2M/T

Where:

- M = number of matches (characters/elements in common subsequences)
- T = total number of elements in both sequences

So, ratio = 1.0 → perfect match, ratio = 0 → no match.

2. Roman to Native Similarity
How it’s computed:
roman_to_native_similarity = calculate_similarity_fixed(roman_to_native_output, native_ground_truth)

- It compares:
  - roman_to_native_output → what your IndicXLit model generated (Roman → Native script conversion).
  - native_ground_truth → the correct Native script from your Excel file.
- Why it matters:
  This checks transliteration accuracy (does the Roman input correctly convert to its native script?).

3. Native to English Similarity
How it’s computed:
native_to_english_similarity = calculate_similarity_fixed(translated_english, english_ground_truth)

- It compares:
  - translated_english → output from IndicTrans2 (Native → English translation).
  - english_ground_truth → the reference English translation from your Excel file.
- Again, difflib.SequenceMatcher is used.
- Why it matters:
  This measures translation quality (semantic + surface similarity) from Native script → English.

---

4. Mapping Columns (for debugging / traceability)

Example Interpretation of a Row

Suppose you had this row:

Language (GT)	IndicLID Conf	LID Mapping Used	Roman→Native Sim	Xlit Used	Native→English Sim	Trans2 Used
Hindi	0.88	hi	92%	hi	87%	hi-en

Interpretation:

- The model was 88% confident the text was Hindi.
- Roman→Native was very accurate (92%).
- Native→English translation was fairly good (87%).
- Pipeline is performing well for this sample.

Limitations of difflib:

- It checks character-level overlap, not meaning.
  E.g., “come fast” vs “hurry up” = low similarity, but semantically correct.

Common Synonyms in Malayalam

Word	Meaning (English)	Synonyms in Malayalam
വലിയ (valiya)	big/large	വലുത് (valuth), വമ്പൻ (vampan), ഭീമൻ (bheeman)
സന്തോഷം (santhosham)	happiness	ആനന്ദം (aanandam), ഹർഷം (harsham), ഉല്ലാസം (ullaasam)
ദുഃഖം (dukkham)	sadness/sorrow	വിഷാദം (vishaadam), വേദന (vedana), വ്യസനം (vyasanam)
ഭക്ഷണം (bhakshanam)	food	കറി (kari), അന്നം (annam), ഉണ്ണുക (unnuka – to eat)

വേഗം (vegam)	speed/fast	ക്ഷിപ്രം (kshipram), ദ്രുതം (drutham), ത്വരം (thvaram)
വീട് (veedu)	house/home	ഗൃഹം (griham), ഭവനം (bhavanam), മന്ദിരം (mandiram)
സ്ത്രീ (stree)	woman	പെൺകുട്ടി (penkutti), വനിത (vanitha), നാരി (naari)
പുരുഷൻ (purushan)	man	ആൺ (aan), മനുഷ്യൻ (manushyan), നരൻ (naran)

- For semantic correctness, you might also consider **BLEU, chrF, or BERTScore** in future.

- IndicLID Confidence = correctness of language detection
- Roman→Native Similarity = transliteration accuracy
- Native→English Similarity = translation accuracy